Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 2, 2025

Single-Person 3D Human Pose Estimation Based on Deep Learning: A Review

Aoyu Xia*, Shouming Hou, Zixuan Lu, Weibo Yang

College of Henan Polytechnic University, Jiaozuo, China
*Corresponding Author: Aoyu Xia

Abstract: With the rapid development of deep learning technologies, human pose estimation has become a hot research topic in the field of computer vision. This paper provides a systematic review of the latest advances in single-person 3D human pose estimation based on deep learning, with a focus on the main challenges faced by both single-view and multi-view approaches, such as pose estimation accuracy, occlusion issues, and the effective utilization of depth information. Firstly, the paper explores single-view human pose estimation methods based on single-frame images and video data. Then, it introduces the research status of multi-view human pose estimation, and discusses in detail how these methods address pose occlusion and multi-view data fusion. In addition, the paper reviews commonly used datasets and evaluation metrics, and assesses the performance of various methods on standard datasets through comparative analysis. Finally, the paper outlines future research directions for single-person 3D human pose estimation, particularly in improving estimation accuracy and addressing pose variations and occlusions in complex scenarios.

Keywords: Deep Learning; Single-Person 3D Human Pose Estimation; Single-View; Multi-View.

1. Introduction

Single-person 3D human pose estimation is a core task in computer vision, aiming to accurately recover the human body's pose in three-dimensional space from images or videos. Unlike traditional 2D pose estimation, 3D pose estimation not only provides the 2D coordinates of joints but also precisely captures depth information and the spatial positioning of the body. Due to its ability to comprehensively depict the spatial relationships of the human body, 3D pose estimation has broad application value in multiple fields, including virtual reality (VR), augmented reality (AR), human-computer interaction, intelligent surveillance, sports analysis, and medical rehabilitation. Accurate 3D pose estimation can enhance the level of system intelligence, improve user experience, and provide technical support and application solutions for related fields.

With the rapid development of deep learning technologies, deep learning-based methods have become the mainstream solution for single-person 3D human pose estimation. Despite significant progress in this field, factors such as pose occlusion, viewpoint variation, and the diversity of human actions still present significant challenges to accurate and efficient 3D pose estimation.

This paper aims to review the latest research developments in single-person 3D human pose estimation based on deep learning. Firstly, the paper will analyze the core issues and challenges in single-view and multi-view pose estimation and explore how existing methods address these challenges. Next, the paper will introduce the main methods based on single-view and multi-view images, focusing on how to effectively fuse depth information and solve pose occlusion issues. Subsequently, the paper will review commonly used datasets and evaluation metrics, assessing the performance of different methods on standard datasets through comparative analysis. Finally, the paper will outline future research directions for single-person 3D human pose estimation, particularly the potential for improving estimation accuracy and addressing challenges in complex scenarios.

The structure of the paper is as follows: the first section outlines the problems and challenges faced in 3D human pose estimation; the second section introduces commonly used datasets and evaluation metrics; the third section reviews the research progress in single-view human pose estimation, discussing methods based on single-frame images and videos; the fourth section discusses relevant research in multi-view human pose estimation; the fifth section reviews existing datasets and evaluation standards; finally, the sixth section summarizes the main conclusions of the paper and outlines future research directions.

2. ISSUES and Challenges

The 3D human pose estimation task faces several challenges. First, pose estimation needs to be performed in complex backgrounds, varying lighting conditions, and dynamic poses. Particularly in severe occlusion or extreme viewpoints, traditional methods often struggle to provide stable and accurate estimations. Secondly, the high degrees of freedom of human joints in three-dimensional space make it more difficult to accurately locate each joint and its relative positions. Since 2D images lack depth information, effectively recovering high-quality 3D poses from 2D images remains a key challenge in current research.

Traditional 3D human pose estimation methods primarily rely on geometric models and rule-based inference, often requiring manual feature design and pose recovery through predefined models. While these methods can be effective in specific scenarios, they have obvious limitations. For instance, they are more sensitive to issues such as pose occlusion and viewpoint changes, leading to large estimation errors. Additionally, traditional methods often require complex preprocessing steps and struggle to adapt to dynamic scenes, making it difficult to handle the diversity and complexity of real-world environments.

In contrast, deep learning methods have brought significant

breakthroughs to 3D human pose estimation. Through end-to-end training, deep learning can automatically extract effective features from large datasets, avoiding the cumbersome manual design process inherent in traditional methods. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Graph Convolutional Networks (GCNs), have demonstrated powerful capabilities in handling complex image data and human poses. These methods can directly learn depth information from single-view or multi-view images and use it to recover 3D poses, greatly improving both accuracy and robustness. Furthermore, deep learning methods effectively mitigate the impact of occlusion and viewpoint changes on estimation results through strategies such as multi-level feature fusion and data augmentation.

3. DATASETS and Evaluation Metrics

3.1. Datasets

Human3.6M: Human3.6M [1] is one of the most widely used datasets for 3D human pose estimation, containing data from 11 participants performing 15 daily activities (such as walking, sitting, eating, etc.), with a total of 3.6 million high-resolution image frames. The dataset not only provides 2D keypoint annotations for each frame of the image but also includes the corresponding 3D skeleton coordinates. The major advantage of Human3.6M is its high-quality annotation data, where all 3D joint positions are obtained using a motion capture system, ensuring high annotation accuracy. Moreover, the dataset covers a wide range of daily activities, making it suitable for tasks such as pose estimation and action recognition. As a key benchmark dataset in 3D human pose estimation, Human3.6M has become a primary reference for model evaluation and research.

Human Eva: Human Eva [2] is a classic dataset designed specifically for multi-view 3D human pose estimation. It captures the actions of three participants from four cameras positioned at different angles, annotating both 3D skeleton positions and the corresponding 2D joint positions. A notable feature of this dataset is its multi-view structure, which helps models handle image data from different angles, significantly improving the robustness of pose estimation, especially when dealing with viewpoint changes and occlusions. Although the types of actions in the Human Eva dataset are relatively simple, mainly including walking, running, and standing, it remains a standard dataset for evaluating the accuracy of multi-view pose estimation algorithms. Table 2 presents the performance of various pose estimation methods on the Human Eva dataset.

MPI-INF-3DHP: The MPI-INF-3DHP [3] dataset, released by the Max Planck Institute in Germany, is designed to provide data for 3D human pose estimation in complex scenarios. Unlike other datasets, MPI-INF-3DHP not only provides multi-view images and accurate 3D joint annotations but also includes more complex action types, such as social interactions and indoor activities. Given that this dataset involves more occlusions, viewpoint variations, and human interactions, it is particularly suitable for evaluating pose estimation and behavior analysis in multi-person scenarios. Table 3 presents the performance of various pose estimation methods on the MPI-INF-3DHP dataset.

3.2. Evaluation Metrics

MPJPE (Mean Per Joint Position Error): MPJPE is a commonly used metric for evaluating the accuracy of 3D pose estimation, calculated by measuring the average Euclidean distance between the predicted 3D joint positions and the ground truth joint positions. The lower the MPJPE, the more accurate the model's pose predictions. This metric is widely used for evaluating models in both single-person and multi-person pose estimation tasks.

P-MPJPE (Procrustes-Aligned MPJPE): P-MPJPE is an improved version of MPJPE, which calculates the error after applying Procrustes alignment to remove rotation, translation, and scale differences in the pose. The advantage of this method is that it eliminates the impact of pose transformations on the evaluation, allowing the focus to be on the relative error in joint positions. P-MPJPE has become a standard evaluation metric in the field of 3D pose estimation.

AUC (Area Under Curve): AUC is used to evaluate a model's performance at different error thresholds by calculating the area under the precision-recall curve. The AUC value ranges between 0 and 1, with a higher value indicating better performance across various prediction error thresholds. AUC is commonly used for multi-task evaluations, especially in multi-class and multi-object scenarios in pose estimation.

PCK (Percentage of Correct Keypoints): PCK is another commonly used evaluation metric for pose estimation, which measures the performance of the model by calculating the error in predicting keypoints compared to the true keypoints. A threshold for error is typically set, and if the predicted joint's error is below the threshold, the joint is considered correctly estimated. PCK is often used to evaluate the robustness of pose estimation models under different error thresholds.

PCP (Percentage of Correct Parts): PCP evaluates the accuracy of the pose estimation for different parts of the body, usually assessing the accuracy of skeletal connections (such as the arms and legs). PCP measures how well the estimated parts align with the true data, and it is commonly used in action recognition and human analysis tasks.

4. Single-View 3D HUMAN Pose Estimation Methods

This section discusses methods for 3D human pose estimation from single-frame images and videos, highlighting the various techniques developed for improving pose estimation accuracy and robustness under different conditions.

4.1. Single-Frame Image-Based Methods

Sun et al. [4] proposed a structure-aware regression method that uses a reparameterized pose representation where bones replace joints. They defined a composite loss function to encode long-range dependencies between joints, effectively utilizing the structural information in the pose. Pavlakos et al. [5] divided the 3D space into multiple small voxels and predicted the probability distribution of joints within each voxel, allowing for a more precise representation of the human body’s spatial structure. The following year, Pavlakos et al. [6] proposed a method that does not require high-precision 3D annotated data. They used the depth-order relationship of human joints for 3D pose estimation, utilizing depth information as a constraint to help predict the joint positions in 3D space. The innovation of this method lies in reducing the reliance on high-quality annotated data, solving the data acquisition problem.

Martinez et al. [7] demonstrated that 3D pose estimation

errors arise from visual analysis of the image and proposed a simple, fast, and lightweight neural network that directly estimates 3D pose from 2D ground truth, achieving state-of-the-art results. Katircioglu et al. [8] noted that many modern methods rely on deep learning, either directly using convolutional neural networks (CNN) to regress 3D poses from images, ignoring dependencies between joints, or using structured learning frameworks to model these dependencies, which leads to high computational costs during inference. To address this, they proposed a deep learning regression architecture that can structurally predict 3D human pose from monocular images or 2D joint location heatmaps. They combined this with an overcomplete autoencoder to learn latent high-dimensional pose representations and introduced an efficient long short-term memory (LSTM) network to enhance temporal consistency in 3D pose prediction.

Li and Lee [9] believed that estimating 3D human pose from monocular input is an inverse problem that may have multiple feasible solutions. To resolve this ambiguity, they proposed a method to generate multiple 3D pose hypotheses, exploring the idea of multi-hypothesis generation to alleviate the ambiguity inherent in traditional methods. Wandt and Rosenhahn [10] introduced a novel network architecture—RepNet (Adversarial Reprojection Network)—that utilizes a generative adversarial network (GAN) framework to enhance 3D pose estimation quality through reprojection. This method overcomes the traditional methods' heavy dependence on high-quality data annotation through adversarial training and weakly supervised learning.

Choi et al. [11] proposed the Pose2Mesh model, which processes the human skeleton structure using graph convolutional networks and directly recovers 3D poses from 2D human keypoints, avoiding the need for 3D annotated data. Additionally, Pose2Mesh can generate high-quality 3D human meshes, widely applied in animation production and virtual reality scenarios. Liu et al. [12] proposed a feature-enhancement network to estimate 3D hand and body poses from a single RGB image. This method enhances the features learned by the convolutional layers through a new long short-term dependency (LSTD) module, enabling the feature map to perceive long- and short-term dependencies between different body parts. Additionally, Liu et al. introduced a context consistency gate (CCG) that modulates the feature map based on the consistency with the context representation, further enhancing the performance of the LSTD module.

Kundu et al. [13] proposed a new unsupervised 3D pose estimation framework that does not require any paired or unpaired weak supervision constraints and relies on minimal prior knowledge. The model defines latent kinematic 3D structures, such as bone length ratios and skeletal joint connection information. It uses three consecutive differentiable transformations, including forward kinematics, camera projection, and spatial map transformation. This design effectively decouples pose representations and generates interpretable latent pose representations without explicit training of a pose mapper. Since it does not rely on unstable adversarial setups, the model can learn from real-world video data, not just laboratory environment data.

Zhou et al. [14] conducted an in-depth analysis of the poor generalization and effectiveness of image features and proposed an improved framework. The framework first selects useful image features through keypoint querying, uses a pose-guided transformation layer to adaptively restrict the update of insignificant image blocks, and then prunes non-significant image blocks from the feature map through an adaptive feature selection module, emphasizing the retained key image features. This progressive learning method effectively avoids further training on irrelevant features.

Table 1 shows the performance comparison of these methods across various datasets.

Table 1. Performance comparison of methods across different datasets, where ' ↓ ' indicates that a lower number is better, ' ↑ ' indicates that a higher number is better, and '-' indicates that the corresponding result was not reported.

	Human3.6M		Human Eva	MPI-INF-3DHP	
Method	MPJPE(mm) ↓	P-MPJPE(mm) ↓	MPJPE(mm) ↓	PCK(%) ↑	AUC(%) ↑
Sun et al. [4]	92.4	59.1	-	-	-
Pavlakos et al. [5]	71.9	-	24.3	-	-
Martinez et al. [7]	62.9	-	24.6	-	-
Pavlakos et al. [6]	56.2	-	18.3	71.9	35.3
Katircioglu et al. [8]	67.3	-	35.6	-	-
Li et al. [9]	52.7	42.6	-	67.9	-
Wandt et al. [10]	89.9	65.1	-	82.5	58.5
Choi et al. [11]	64.9	-	-	-	-
Liu et al. [12]	61.1	-	-	69.6	-
Kundu et al. [13]	56.1	-	-	81.9	52.6
Zhou et al. [14]	51.0	-	-	88.2	59.3

4.2. Video-based Methods

Video data provides rich spatiotemporal contextual information, which can significantly improve the accuracy and robustness of 3D human pose estimation (3D HPE). Tekin et al. [15] proposed a 3D pose recovery method that utilizes motion information from consecutive frames in a video sequence. This method regresses the 3D pose in the spatiotemporal volume and compensates for motion across consecutive frames, thereby keeping the subject within the central frame and overcoming the ambiguity problem. Pavllo et al. [16] demonstrated that 3D poses in videos can be effectively estimated using an expanded temporal convolutional fully convolutional model based on 2D keypoints. This method uses semi-supervised training, where 2D keypoints are first predicted for unlabeled videos, followed by 3D pose estimation, and ultimately, the 2D keypoints are reprojected back.

Zhang et al. [17] proposed a new dynamic graph network (DG-Net), which can dynamically recognize the affinity of human body joints and enhance the accuracy of 3D pose estimation by adaptively learning spatial and temporal joint relationships in the video. Unlike traditional graph convolution methods, Zhang et al. introduced dynamic spatiotemporal graph convolution (DSG/DTG), which discovers joint affinities based on spatial distance and

temporal motion similarity, thus reducing the impact of depth ambiguity and motion uncertainty on 3D pose improvement.

Inspired by Newell et al. [19]'s stacked hourglass network, Xu and Takano [18] proposed a graph convolutional network architecture called Graph Stacked Hourglass Network for 2D-to-3D human pose estimation tasks. The network consists of multiple encoder-decoder modules that can process graph-structured features across three different scales of human skeletal representations, learning both local and global feature representations. Zheng et al. [20] proposed a purely Transformer-based method called PoseFormer for 3D human pose estimation in videos. This method discards the traditional convolutional structures and designs a spatiotemporal Transformer architecture to comprehensively model the joint relationships within each frame and the temporal correlations between frames, ultimately outputting the precise 3D pose of the central frame.

In subsequent research, Zhao et al. [21] noticed that PoseFormer incurs exponential computational costs when processing long sequences and is highly sensitive to 2D pose noise. To address this, they proposed PoseFormerV2, which introduces frequency domain representations to process input joint sequences and designs a spatiotemporal feature fusion module to reduce the gap between temporal and frequency domain features, balancing speed and accuracy. Zhang et al. [22] observed the differences in the motion characteristics of different joints and proposed MixSTE (Mixed Spatiotemporal Encoder). This method models the temporal motion of each joint individually using temporal Transformer blocks, learns spatial correlations between joints using spatial Transformer blocks, and alternates between these blocks to obtain more effective spatiotemporal feature encoding.

Gong et al. [23] introduced a diffusion model, treating 3D pose estimation as a reverse diffusion process. Through specific pose initialization, a Gaussian mixture model's forward diffusion process, and a reverse diffusion process with contextual conditions, this model can naturally handle the uncertainty and ambiguity of 3D poses. Tang et al. [24] noted that Transformer-based models incur high computational costs when calculating affinity matrices for video sequences. They proposed a spatiotemporal cross-attention (STC) block, which divides the input features evenly along the channel dimension into two partitions. Spatial and temporal attention are then performed on each partition, and by concatenating the outputs of the attention layers, the model can simultaneously model interactions between joints in the same frame and on the same trajectory.

Zhong et al. [25] pointed out that existing video-based methods typically split long video sequences into multiple short subsequences for separate processing, which leads to the loss of complementary information between sequences. Therefore, they proposed a frame padding multi-scale transformation method, which includes a frame padding step in video sequence preprocessing and a multi-scale temporal transformation backbone. By padding and enhancing the input frames, this network can better understand and process feature information from different time steps, thus solving the information loss problem.

Table 2 shows the performance comparison of these methods across various datasets.

Table 2. Performance comparison of methods across different datasets, where ' ↓ ' indicates that a lower number is better, ' ↑ ' indicates that a higher number is better, and '-' indicates that the corresponding result was not reported.

Method	Human3.6M		Human Eva-I	MPI-INF-3DHP	
	MPJPE(mm) ↓	P-MPJPE(mm) ↓	P-MPJPE(mm) ↓	PCK(%) ↑	AUC(%) ↑
Tekin et al. [15]	124.97	-	56.6	-	-
Pavllo et al. [16]	46.8	36.5	23.1	-	-
Zhang et al. [17]	45.3	35.4	19.5	87.5	53.8
Xu et al. [18]	51.9	35.8	-	80.1	45.8
Zheng et al. [20]	44.3	34.6	-	88.6	56.4
Zhang et al. [22]	39.8	30.6	-	94.4	66.5
Gong et al. [23]	36.9	-	-	98.0	75.9
Tang et al. [24]	40.5	31.8	-	98.7	83.9
Zhao et al. [21]	45.2	35.6	-	97.9	78.8
Zhong et al. [25]	40.1	32.0	-	98.6	71.5

5. Multi-View 3D HUMAN Pose Estimation Methods

Multi-view 3D human pose estimation methods integrate image data from multiple camera views, utilizing information from different angles to accurately estimate the 3D pose of the human body. Compared to single-view methods, multi-view methods effectively address occlusion, viewpoint limitations, and depth estimation issues, leading to higher precision in 3D pose prediction.

Iskokav et al. [26] proposed two multi-view 3D human pose estimation methods based on learnable triangulation. The first method is a differentiable algebraic triangulation approach that estimates 3D pose by utilizing weighted confidence predictions from input images. The second method aggregates volume from intermediate 2D feature maps and refines the aggregated volume using 3D convolutions to generate 3D joint heatmaps while allowing for modeling the prior of human poses. Remelli et al. [27] introduced a lightweight solution that recovers 3D poses from multi-view images captured by spatially calibrated cameras. This method integrates input images into a unified latent pose representation using 3D geometry, decouples from camera viewpoints, and avoids the use of computationally intensive volumetric grids, enabling efficient 3D pose inference from different perspectives.

Kim et al. [28] proposed a self-supervised training method that learns relative depths between joints from unlabeled multi-view videos without camera calibration, satisfying multi-view consistency constraints. This method improves network performance by designing a multi-hypothesis regularized network (MHCanonNet) to accurately estimate the final 3D pose through diversified feature extraction. Liang

and Lin et al. [29] introduced a new framework that simultaneously estimates the pose and shape of the human body from multi-view images. They introduced a shape-aware mechanism that combines pose optimization and shape estimation, thus enabling more precise and detailed 3D human reconstruction without relying on accurate 3D data.

Chen et al. [30] proposed an efficient cross-view tracking method that combines multi-view image data to perform keypoint matching and tracking across different views, using spatial consistency constraints to ensure consistency of keypoints in 3D space. This method models temporal continuity to maintain the stability of human poses and can perform high-precision multi-person 3D pose estimation at ultra-high frame rates (over 100 FPS). Xie et al. [31] proposed the MetaFuse model, which adopts a pre-trained model strategy and learns more generalized and accurate feature representations by pre-training on a large-scale 3D human pose dataset. By integrating multi-modal information such as 2D and 3D keypoints, it provides a comprehensive description of human poses.

Zhang et al. [32] proposed an innovative method called 4D Association Graph, aimed at real-time multi-person motion capture using multiple cameras. This method constructs a four-dimensional association graph that fuses time (temporal sequences), space (multi-view), and human skeleton information (pose) to enable real-time estimation and tracking of multi-person poses. With this design, the model can efficiently associate data across multiple views and accurately capture and synchronize motion information from different cameras.

Dong et al. [33] proposed a Shape-Aware multi-person pose estimation method that uses multi-view images for pose estimation and enhances accuracy with a shape-aware mechanism. In multi-view scenarios, the model can accurately estimate occluded or partially invisible keypoints by learning and utilizing human shape features, improving the robustness of multi-person pose estimation, especially in complex scenes and during human interaction.

Ye et al. [34] proposed Faster VoxelPose, a real-time 3D human pose estimation method based on orthogonal projections. By converting the 3D pose estimation problem into multiple 2D view projection problems, this method significantly enhances computational efficiency. It can process multi-person 3D pose estimation in real-time scenarios and improves 3D keypoint prediction accuracy with an enhanced convolutional neural network architecture.

Dong et al. [35] introduced a fast and robust multi-person 3D pose estimation and tracking method designed specifically for complex scenes with multiple viewpoints. By combining multi-view image information and advanced optimization strategies, this method achieves efficient 3D pose estimation and real-time tracking. Its innovation lies in the introduction of a new multi-view data fusion mechanism, allowing pose estimation and tracking across multiple viewpoints, overcoming pose estimation inaccuracies caused by insufficient viewpoints or occlusion in traditional methods.

Table 3 presents a performance comparison of these methods across various datasets.

Table 3. Performance comparison of methods across different datasets, where ' ↓ ' indicates that a lower number is better, ' ↑ ' indicates that a higher number is better, and '-' indicates that the corresponding result was not reported.

	Human3.6M		**Total Capture**	**Campus**	**Shelf**
Method	MPJPE(mm) ↓	P-MPJPE(mm) ↓	MPJPE(mm) ↓	PCP(%) ↑	PCP(%) ↑
Iskakov et al. [26]	20.8	-	-	-	-
Liang et al. [29]	44.4	-	-	-	-
Chen et al. [30]	-	-	-	96.6	96.8
Remelli et al. [27]	30.2	-	-	-	-
Xie et al. [31]	29.3	-	32.4	-	-
Zhang et al. [32]	-	-	-	-	97.6
Dong et al. [33]	-	-	-	-	96.9
Ye et al. [34]	-	-	-	96.2	97.6
Dong et al. [35]	-	-	-	96.3	96.9
Kim et al. [28]	69.0	50.3	-	-	-

6. Conclusion

This paper provides a comprehensive analysis of single-person 3D human pose estimation based on deep learning, discussing in detail the latest advances in single-view and multi-view methods, as well as the main challenges they face. The paper summarizes how different methods address issues such as accuracy, occlusion, complex backgrounds, and viewpoint changes in pose estimation, highlighting the key role of deep learning techniques in improving estimation accuracy and robustness. Finally, the paper lists commonly used human pose estimation datasets and evaluation metrics, and compares the performance of various methods across different datasets.

Future development of 3D human pose estimation will focus on improving model accuracy and robustness in complex scenarios, particularly under conditions of dynamic changes, occlusion, and varying lighting. Meanwhile, enhancing real-time performance will become a key direction, especially in applications such as virtual reality (VR) and augmented reality (AR), which require high real-time feedback. In addition, automated data annotation methods and richer datasets will help address the challenges of data scarcity and annotation difficulties, promoting the widespread application of these models. Overall, future research will aim to balance accuracy, efficiency, and applicability to meet the needs of different industries.

7. CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

Acknowledgements

The research leading to these results received funding from Scientific and Technological Project of Henan Province under Grant Agreement No.232102320171 and Soft Science Research Project of Henan Province under Grant Agreement No. 242400410187. We would like to thank the Henan

Provincial Government for their financial support of this research.

References

[1] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, Jul. 2014, doi: 10.1109/TPAMI.2013.248.

[2] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," Int J Comput Vis, vol. 87, no. 1–2, pp. 4–27, Mar. 2010, doi: 10.1007/s11263-009-0273-6.

[3] D. Mehta et al., "Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision," in 2017 International Conference on 3D Vision (3DV), Qingdao: IEEE, Oct. 2017, pp. 506–516. doi: 10.1109/3DV.2017.00064.

[4] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional Human Pose Regression".

[5] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 1263–1272. doi: 10.1109/CVPR.2017.139.

[6] G. Pavlakos, X. Zhou, and K. Daniilidis, "Ordinal Depth Supervision for 3D Human Pose Estimation," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT: IEEE, Jun. 2018, pp. 7307–7316. doi: 10.1109/CVPR.2018.00763.

[7] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A Simple Yet Effective Baseline for 3d Human Pose Estimation," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, Oct. 2017, pp. 2659–2668. doi: 10.1109/ICCV.2017.288.

[8] I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, and P. Fua, "Learning Latent Representations of 3D Human Pose with Deep Neural Networks," Int J Comput Vis, vol. 126, no. 12, pp. 1326–1341, Dec. 2018, doi: 10.1007/s11263-018-1066-6.

[9] C. Li and G. H. Lee, "Generating Multiple Hypotheses for 3D Human Pose Estimation With Mixture Density Network," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 9879–9887. doi: 10.1109/CVPR.2019.01012.

[10] B. Wandt and B. Rosenhahn, "RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7774–7783. doi: 10.1109/CVPR.2019.00797.

[11] H. Choi, G. Moon, and K. M. Lee, "Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose," in Computer Vision – ECCV 2020, Springer, Cham, 2020, pp. 769–787. doi: 10.1007/978-3-030-58571-6_45.

[12] J. Liu et al., "Feature Boosting Network For 3D Pose Estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 494–501, Feb. 2020, doi: 10.1109/TPAMI.2019.2894422.

[13] J. N. Kundu, S. Seth, R. M V, M. Rakesh, V. B. Radhakrishnan, and A. Chakraborty, "Kinematic-Structure-Preserved Representation for Unsupervised 3D Human Pose Estimation," AAAI, vol. 34, no. 07, pp. 11312–11319, Apr. 2020, doi: 10.1609/aaai.v34i07.6792.

[14] F. Zhou, J. Yin, and P. Li, "Lifting by Image–Leveraging Image Cues for Accurate 3D Human Pose Estimation," in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 7632–7640. Accessed: Apr. 15, 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28596

[15] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, "Direct Prediction of 3D Body Poses from Motion Compensated Sequences," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 991–1000. doi: 10.1109/CVPR.2016.113.

[16] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7745–7754. doi: 10.1109/CVPR.2019.00794.

[17] J. Zhang, Y. Wang, Z. Zhou, T. Luan, Z. Wang, and Y. Qiao, "Learning Dynamical Human-Joint Affinity for 3D Pose Estimation in Videos," IEEE Trans. on Image Process., vol. 30, pp. 7914–7925, 2021, doi: 10.1109/TIP.2021.3109517.

[18] T. Xu and W. Takano, "Graph Stacked Hourglass Networks for 3D Human Pose Estimation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, Jun. 2021, pp. 16100–16109. doi: 10.1109/CVPR46437.2021.01584.

[19] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," European Conference on Computer Vision, 2016, doi: 10.1007/978-3-319-46484-8_29.

[20] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D Human Pose Estimation with Spatial and Temporal Transformers," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 11636–11645. doi: 10.1109/ICCV48922.2021.01145.

[21] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 8877–8886. doi: 10.1109/CVPR52729.2023.00857.

[22] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, Jun. 2022, pp. 13222–13232. doi: 10.1109/CVPR52688.2022.01288.

[23] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "DiffPose: Toward More Reliable 3D Pose Estimation," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 13041–13051. doi: 10.1109/CVPR52729.2023.01253.

[24] Z. Tang, Z. Qiu, Y. Hao, R. Hong, and T. Yao, "3D Human Pose Estimation with Spatio-Temporal Criss-Cross Attention," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 4790–4799. doi: 10.1109/CVPR52729.2023.00464.

[25] Y. Zhong, G. Yang, D. Zhong, X. Yang, and S. Wang, "Frame-Padded Multiscale Transformer for Monocular 3D Human Pose Estimation," IEEE Trans. Multimedia, vol. 26, pp. 6191–6201, 2024, doi: 10.1109/TMM.2023.3347095.

[26] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable Triangulation of Human Pose," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul,

Korea (South): IEEE, Oct. 2019, pp. 7717–7726. doi: 10.1109/ICCV.2019.00781.

[27] E. Remelli, S. Han, S. Honari, P. Fua, and R. Wang, "Lightweight Multi-View 3D Pose Estimation Through Camera-Disentangled Representation," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2020, pp. 6039–6048. doi: 10.1109/CVPR42600.2020.00608.

[28] H.-W. Kim et al., "MHCanonNet: Multi-Hypothesis Canonical lifting Network for self-supervised 3D human pose estimation in the wild video," Pattern Recogn, vol. 145, p. 109908, Jan. 2024, doi: 10.1016/j.patcog.2023.109908.

[29] J. Liang and M. Lin, "Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 4351–4361. doi: 10.1109/ICCV.2019.00445.

[30] L. Chen, H. Ai, R. Chen, Z. Zhuang, and S. Liu, "Cross-View Tracking for Multi-Human 3D Pose Estimation at Over 100 FPS," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2020, pp. 3276–3285. doi: 10.1109/CVPR42600.2020.00334.

[31] R. Xie, C. Wang, and Y. Wang, "MetaFuse: A Pre-trained Fusion Model for Human Pose Estimation," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2020, pp. 13683–13692. doi: 10.1109/CVPR42600.2020.01370.

[32] Y. Zhang, L. An, T. Yu, X. Li, K. Li, and Y. Liu, "4D Association Graph for Realtime Multi-Person Motion Capture Using Multiple Video Cameras," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2020, pp. 1321–1330. doi: 10.1109/CVPR42600.2020.00140.

[33] Z. Dong, J. Song, X. Chen, C. Guo, and O. Hilliges, "Shape-aware Multi-Person Pose Estimation from Multi-View Images," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 11138–11148. doi: 10.1109/ICCV48922.2021.01097.

[34] H. Ye, W. Zhu, C. Wang, R. Wu, and Y. Wang, "Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection," in Computer Vision – ECCV 2022, vol. 13666, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13666. , Cham: Springer Nature Switzerland, 2022, pp. 142–159. doi: 10.1007/978-3-031-20068-7_9.

[35] J. Dong et al., "Fast and Robust Multi-Person 3D Pose Estimation and Tracking From Multiple Views," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 6981–6992, Oct. 2022, doi: 10.1109/TPAMI.2021.3098052.