iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76081

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

3D Human Motion Reconstruction from 2D Image Using Deep Learning and Computer Vision

Dr. S Gunasekaran[1], Archana M[2], Mihikha S[3], Minikha S[4], Upanya K[5]
[1]*Professor in CSE, Ahalia School of Engineering and Technology, Palakkad, Kerala*
[2, 3, 4, 5]*Ahalia School of Engineering and Technology, Palakkad,Kerala*

Abstract: ***In recent years, the field of computer vision has witnessed significant advancements in understanding and interpreting human motion through pose estimation techniques. While traditional 2D human pose estimation methods are capable of detecting body joints from images, they fail to capture the depth, structure, and realistic appearance of the human body. To overcome this limitation, this project focuses on the development of a system for 3D Human Pose Estimation with Realistic 3D Output, which reconstructs a lifelike three-dimensional human model from a single 2D image or video frame using deep learning techniques. The proposed system follows a structured pipeline that includes 2D keypoint detection, 3D pose estimation, and 3D mesh reconstruction to achieve accurate and visually realistic results.The system employs pretrained deep learning models such as HMR (Human Mesh Recovery), SPIN (SMPLify-in-the-Loop), and PARE (Part Attention Regressor) to predict 3D joint coordinates and human body shape parameters efficiently. These parameters are then passed to the SMPL (Skinned Multi-Person Linear) parametric model to generate a smooth and anatomically correct 3D human mesh. The reconstructed model is rendered and visualized using advanced 3D rendering tools, allowing users to rotate, zoom, and observe the model from different viewpoints. The performance of the system is evaluated using standard accuracy metrics such as MPJPE and PA-MPJPE, ensuring reliable and precise pose estimation. This project demonstrates that realistic 3D human reconstruction can be achieved without complex motion capture systems or expensive hardware. By effectively bridging the gap between 2D image perception and 3D geometric understanding, the proposed system offers a practical and scalable solution for realistic human modeling. The results obtained from this work are highly suitable for real-world applications such as animation, sports analysis, motion tracking, virtual reality, and healthcare monitoring.***

Keywords: 3D Human Pose Estimation, Deep Learning, Human Mesh Reconstruction, SMPL Model, HMR, SPIN, PARE, Computer Vision, Realistic 3D Output, Motion Analysis

I. INTRODUCTION

Human motion and posture play an essential role in the way people communicate, interact, and understand their surroundings. In the field of computer vision, enabling machines to accurately interpret and recreate human movement has long been a challenging yet fascinating goal. Human pose estimation allows computers to detect and track the positions of human body joints from images or videos. Although traditional 2D pose estimation techniques have achieved great success in identifying body keypoints on flat images, they are still limited because they cannot represent depth, body structure, or realistic human shape. Due to this limitation, their use becomes restricted in advanced applications such as animation, virtual reality, sports performance analysis, medical rehabilitation, and human–computer interaction.

To overcome these challenges, researchers have increasingly shifted their focus toward 3D human pose estimation, which aims to reconstruct the full three-dimensional structure of the human body from visual data. With the rapid advancement of deep learning, particularly convolutional and regression-based neural networks, it has become possible to extract meaningful features from images and accurately predict 3D human poses.

Modern 3D pose estimation systems no longer focus only on skeletal joint positions but also aim to generate realistic 3D human body meshes that reflect true body shape, proportions, and posture. This progress has been made possible through parametric human body models such as SMPL, which mathematically represent the human body using pose and shape parameters. By combining such models with powerful pretrained deep learning networks, it is now possible to reconstruct a lifelike digital human model even from a single image.

The main objective of this project, “3D Human Motion Reconstruction From 2D Image Using Deep Learning And Computer Vision ”, is to design and implement a system that can accurately estimate 3D human poses and generate visually realistic 3D human body models using deep learning techniques.

©IJRASET: All Rights are Reserved

The proposed system follows a structured pipeline that begins with detecting 2D body keypoints from input images, followed by estimating 3D joint coordinates and body shape parameters, and finally reconstructing a complete 3D human mesh.
The reconstructed model is then visualized using 3D rendering tools, allowing users to observe the human figure from different angles. This approach removes the dependency on expensive motion-capture setups and specialized equipment, making realistic 3D human modeling more accessible and practical.
Together, the four reviewed papers clearly show the research progression from early 2D pose estimation methods to 3D skeletal reconstruction and finally to realistic 3D human mesh recovery.
These studies collectively analyze benchmark datasets, deep learning architectures, optimization techniques, and evaluation protocols, providing a strong theoretical foundation for this project.
Based on the knowledge gained from these surveys, this project integrates 2D pose detection, 3D pose regression, and SMPL-based mesh reconstruction to generate lifelike 3D human models from ordinary images. By leveraging pretrained models such as HMR, SPIN, and PARE, and by following the evaluation standards discussed in the literature, this project aims to develop a practical, efficient, and visually realistic 3D human pose estimation system.
Thus, the literature survey not only highlights the evolution of research in this field but also directly guides the design, methodology, and evaluation of the proposed system.

II.LITERATURE REVIEW

This section explores four important research papers that explain how human pose estimation has evolved over time. Together, these papers show the journey from simple 2D joint detection to advanced 3D pose and realistic human mesh reconstruction, which forms the foundation for this project.

A. Advances in Human Pose Estimation, Tracking, and Action Recognition: A Comprehensive Survey

The work published by Zhou, L et.al. [1] provides a comprehensive overview of recent advances in 2D and 3D human pose estimation, pose tracking, and action recognition for both single and multi-person scenarios in images and videos. It highlights the evolution from traditional regression-based methods to more accurate heatmap-based approaches, which better capture spatial details. The survey also emphasizes the growing impact of transformer architectures and multi-stage models, which have significantly improved performance.
A key trend discussed is the shift toward end-to-end joint frameworks that combine pose estimation, tracking, and action recognition to reduce error propagation and improve robustness. Major challenges such as occlusion, scale variation, low-resolution inputs, and high computational cost are addressed, with future directions pointing toward unified models, multi-modal fusion, and zero-shot learning.
The reviewed methods are grouped into CNN-, RNN-, GCN-, and Transformer-based approaches. CNNs remain the backbone for extracting visual features in both 2D and 3D pose estimation and action recognition. RNNs are mainly used for modeling temporal dynamics in video-based pose estimation and skeleton-based action recognition.
GCNs effectively exploit the graph structure of the human skeleton, making them especially powerful for 3D pose estimation and skeleton-based action recognition. Transformers further enhance performance by capturing global spatial and temporal relationships, often working in hybrid models with GCNs.
The survey also reviews major benchmark datasets such as COCO, MPII, PoseTrack, Human3.6M, MPI-INF-3DHP, MuPoTS-3D, and NTU RGB+D, which support research across 2D pose estimation, 3D pose estimation, tracking, and action recognition. Rather than proposing a single new algorithm, the paper serves as a broad review of existing techniques.
It covers top-down and bottom-up strategies for both 2D and 3D pose estimation, weakly and self-supervised learning for handling limited 3D data, and the growing use of hybrid GCN–Transformer models for action recognition. Finally, it discusses 2D and 3D pose tracking, noting a recent shift from multi-stage to one-stage, end-to-end tracking frameworks for better robustness in real-world scenarios.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

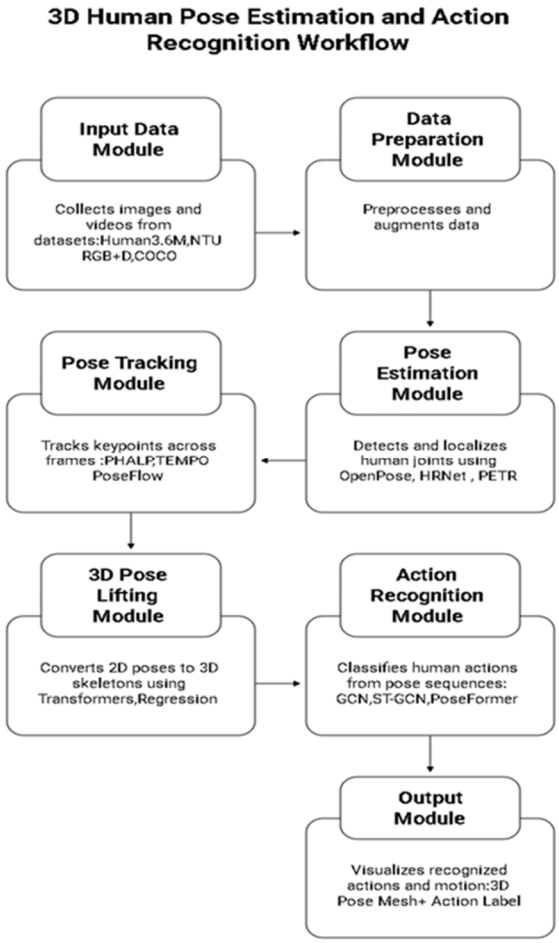


Fig. 1.Workflow of this methodology

The system begins with data preparation using standard 2D and 3D human pose datasets such as Human3.6M and MuPoTS-3D. The input images are normalized, and variations like occlusion and scale changes are handled to improve model robustness.

For pose estimation, a top-down approach is adopted. A person detector such as Mask R-CNN first identifies individuals in each frame. Then, deep networks like HRNet or Transformer-based models are used to estimate 2D keypoints. For 3D pose estimation, a 2D-to-3D lifting network converts these keypoints into 3D joint coordinates, often using temporal information from video.

In the pose tracking module, detected poses are linked across frames using motion and similarity cues to maintain identity consistency. The system also includes a recovery mechanism to handle tracking failures in dynamic scenes.

For action recognition, the tracked pose sequences are fed into temporal models such as RNNs or Transformers. The system mainly relies on skeleton-based features for robustness but can optionally combine appearance or optical flow features to improve accuracy.

Since this work is a survey, it does not report new experimental results but reviews the performance of existing methods using standard evaluation metrics. For 2D pose estimation, metrics such as Average Precision (AP), mAP, and PCK are commonly used. In 3D pose estimation, accuracy is mainly measured using MPJPE and PMPJPE, along with PCK and AUC. Pose tracking performance is evaluated with MOTA and PCP, while action recognition is measured using classification accuracy, often under cross-subject and cross-view settings.

The survey concludes that heatmap-based methods outperform regression-based methods in 2D pose estimation. Top-down approaches generally provide higher accuracy than bottom-up ones for multi-person cases, while video-based 3D methods benefit from temporal information. For action recognition, GCN- and Transformer-based models, especially hybrid approaches, achieve the best performance on skeleton data.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. A Survey on Deep 3D Human Pose Estimation

The work published by Kan Li et al. [2] 3D Human Pose Estimation (3D-HPE) is a key task in computer vision with applications in XR, action recognition, and human-computer interaction. Modern approaches often rely on 2D pose estimation as an intermediate step, categorized into two-stage (2D-to-3D lifting) and single-stage methods. Multi-person 3D-HPE typically uses top-down, bottom-up, or hybrid strategies, while solution spaces include deterministic, probabilistic, and emerging NeRF-based techniques. The field faces challenges like depth ambiguity, occlusion, articulation variability, data scarcity, and computational complexity, especially in monocular setups or unconstrained environments. Architecturally, CNNs, GCNs, Transformers, and hybrid models dominate. CNNs extract spatial features and predict 2D keypoints or 3D coordinates, often combined with temporal models for video sequences. GCNs leverage the human skeleton's graph structure to capture joint relationships and enforce body consistency, effectively addressing occlusion and depth ambiguity. Transformers model long-range spatial and temporal dependencies, excelling in video-based 3D-HPE but requiring large datasets and high computation. Hybrid models combine these architectures to capture local, graph-based, and global relationships simultaneously, improving robustness and accuracy.

Learning paradigms span supervised, weakly-supervised, unsupervised, and self-supervised approaches, with strategies like reconstruction loss, geometric constraints, and adversarial training to handle limited 3D annotations. Methods are also categorized by input modality, including monocular and multi-view images or videos, as well as single-person vs. multi-person scenarios. Advanced techniques include 2D-to-3D lifting, single-stage regression, probabilistic modeling, diffusion-based refinement, and SMPL/NeRF representations for detailed human body and scene modeling. Data augmentation and domain adaptation further enhance robustness and generalization across diverse environments.

Benchmark datasets remain crucial for development and evaluation. Human3.6M and MPI-INF-3DHP provide controlled single-person settings, while MuPoTS-3D targets multi-person scenarios. 3DPW offers real-world outdoor data, and synthetic datasets like SURREAL or multi-view datasets such as CMU Panoptic and Total Capture enable pre-training and multi-person evaluation. These datasets, combined with advanced evaluation metrics, drive progress in accuracy, temporal consistency, and real-world applicability.

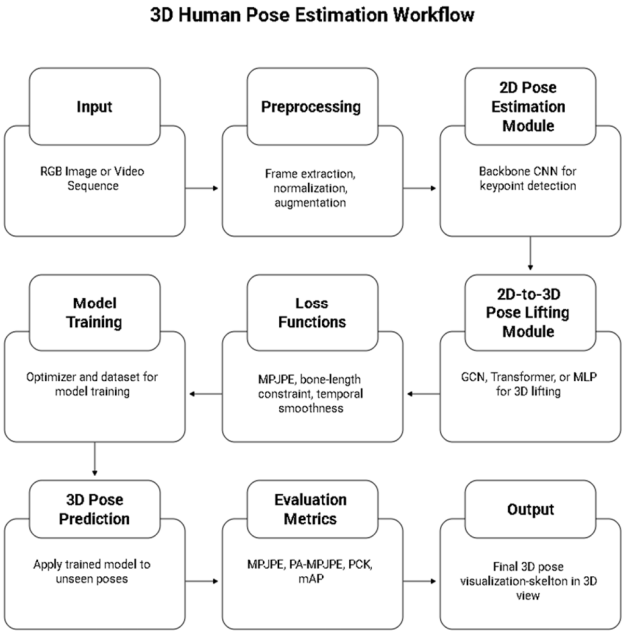


Fig. 2.Workflow of methodology

This study leverages standard 3D human pose datasets for training and evaluation. Human3.6M provides large-scale indoor data with accurate 3D joint annotations captured via multi-camera systems. MPI-INF-3DHP adds indoor and outdoor scenes for better generalization, while MuPoTS-3D focuses on multi-person outdoor poses. 3DPW (3D Poses in the Wild) captures real-world scenarios using IMUs and multi-view cameras. For 2D keypoints, datasets like COCO and MPII are used to pre-train or fine-tune 2D pose detectors.

Most approaches follow a two-stage pipeline, where 2D keypoints are first detected from RGB images using networks like HRNet, Stacked Hourglass, or CPN, and then lifted to 3D coordinates using regression models, GCNs, or Transformer-based networks. Alternative methods include single-stage models that directly regress 3D poses, multi-view fusion for leveraging multiple camera inputs, and self- or weakly-supervised learning to reduce reliance on labeled 3D data. GCNs are particularly effective at modeling joint relationships and maintaining skeletal consistency, while hybrid GCN-Transformer models capture both local and global dependencies, improving robustness in video sequences.

Data preprocessing involves normalizing 2D keypoints and applying augmentations such as rotation, scaling, and occlusion simulation. The 3D pose lifting step uses fully connected networks, GCNs, or Transformers to predict 3D joint coordinates, either person- or camera-centric. Common loss functions include MPJPE (Mean Per Joint Position Error), PA-MPJPE (Procrustes-aligned MPJPE), along with bone-length consistency and temporal smoothness for video. Optimization is typically done using Adam with learning rate scheduling.

Performance is evaluated using metrics like MPJPE, PCK (Percentage of Correct Keypoints), PCP (Percentage of Correct Parts), PSIM (Pose Similarity), and TR loss (Temporal Reconstruction) to assess both accuracy and temporal consistency. Scenario-specific trends show that GCN-based models perform well for single-person 3D-HPE on Human3.6M, GAN-based models excel on MPI-INF-3DHP, and hybrid GCN-Transformer models achieve top results in video settings. Transformers dominate multi-view and multi-person scenarios, particularly in video-based tasks.

Despite these advancements, challenges remain. Limited and less diverse 3D datasets restrict model generalization, and occlusion, lighting variations, background clutter, and motion blur can degrade performance. Two-stage methods also suffer from error propagation from noisy 2D detections, and high computational cost remains a barrier for real-time or mobile applications.

C. Deep Learning -Based Human Pose Estimation: A Survey

The work published by Zheng et.al. [3] Human pose estimation (HPE) is a fundamental task in computer vision that focuses on identifying and localizing human body joints from visual data. The paper “Deep Learning–Based Human Pose Estimation: A Survey” by Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, and Chen Chen provides a comprehensive examination of how modern deep learning techniques have transformed this domain. Human pose estimation serves as the foundation for numerous applications such as human–computer interaction, healthcare monitoring, sports analysis, animation, surveillance, and augmented or virtual reality, making it a critical component in systems that aim to interpret or analyse human behaviour.

Zheng et al. [3] categorize 2D human pose estimation methods into regression-based and detection-based approaches. Regression-based methods, such as DeepPose, directly predict the coordinates of body keypoints from images. Although simple, they often struggle to preserve spatial relationships between joints and are sensitive to occlusions. In contrast, detection-based or heatmap-based methods predict spatial confidence maps for each joint, allowing for more accurate localization. Popular examples include Convolutional Pose Machines (CPM) and Stacked Hourglass Networks, which model spatial dependencies across multiple scales. Later architectures like the High-Resolution Network (HRNet) maintain high-resolution feature maps throughout the network and achieve state-of-the-art accuracy in 2D benchmarks.

For multi-person pose estimation, the authors distinguish between top-down and bottom-up frameworks. Top-down methods first detect individual persons and then apply single-person pose estimation within each detected region. This approach yields high accuracy but requires running separate pose networks for each detected person, increasing computational cost. On the other hand, bottom-up methods such as OpenPose and Part Affinity Fields (PAFs) detect all keypoints in the image simultaneously and then group them into individuals based on learned association maps. These approaches are faster and suitable for real-time applications but may be less precise when handling crowded or overlapping scenes.

In 3D human pose estimation, Zheng et al. classify the existing literature into model-free and model-based methods. Model-free methods either predict 3D joint coordinates directly from RGB images or use 2D-to-3D lifting techniques that reconstruct 3D poses from previously estimated 2D keypoints. The well-known Martinez baseline is an example of a simple yet effective 2D-to-3D lifting model that employs fully connected layers to infer 3D joint positions. In contrast, model-based approaches incorporate prior knowledge about human anatomy by utilizing parametric models such as SMPL (Skinned Multi-Person Linear Model), SMPL-X, and GHUM. Methods like Human Mesh Recovery (HMR), SPIN, and VIBE regress the parameters of these body models to generate full 3D meshes with realistic body shape and posture. Although model-based methods produce more detailed and physically plausible results, they often face difficulties when applied to in-the-wild datasets due to limited 3D annotations, complex backgrounds, and occlusions.

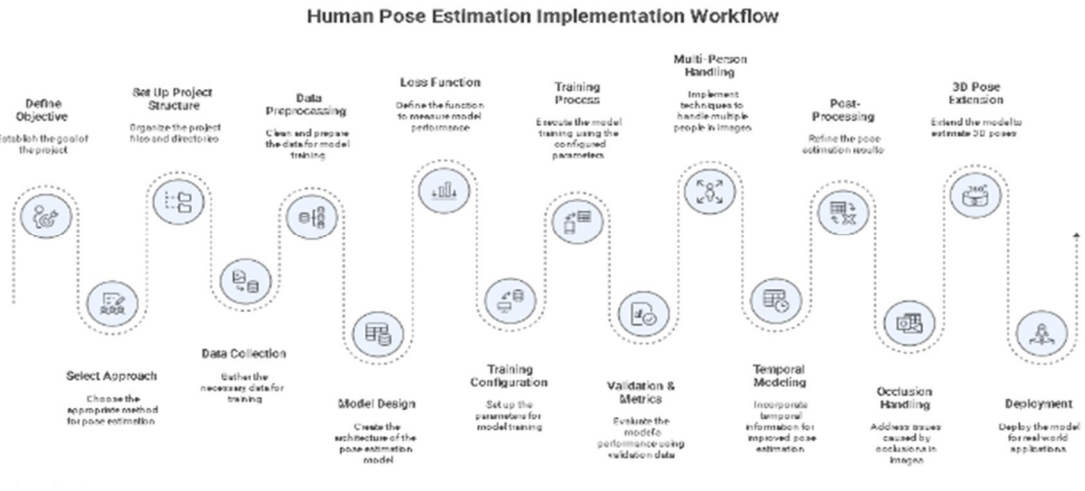


Fig. 3.Workflow of methodology

The survey also presents a comprehensive analysis of commonly used datasets and evaluation metrics. For 2D pose estimation, widely used datasets include MPII, COCO, LSP, FLIC, and PoseTrack, while the most popular metrics are Percentage of Correct Keypoints (PCK), PCKh (head-normalized PCK), mean Average Precision (mAP), and Object Keypoint Similarity (OKS). For 3D pose estimation, the major datasets are Human3.6M, MPI-INF-3DHP, 3DPW, AMASS, and SURREAL. The evaluation metrics for 3D HPE include Mean Per Joint Position Error (MPJPE) and PA-MPJPE, which measures joint prediction error after alignment. Zheng and co-authors note that most 3D datasets are captured in controlled indoor environments, which limits the generalization of trained models to real-world outdoor scenarios.

In addition to pose estimation, the authors review recent progress in pose tracking and action recognition, where temporal modeling plays a crucial role. Techniques such as PoseFlow and PHALP integrate temporal consistency to improve tracking across video frames. Sequential models like Long Short-Term Memory (LSTM) networks and Temporal Convolutional Networks (TCN) capture temporal dependencies to recognize human activities more effectively. Recently, transformer-based models such as PoseFormer have shown promising results by capturing both spatial and temporal relationships simultaneously, leading to more accurate and stable predictions for human motion understanding.

Zheng et al. [3] also compare the performance of existing methods across popular benchmarks. Their analysis reveals that heatmap-based networks combined with strong convolutional backbones, such as HRNet or Stacked Hourglass, consistently outperform direct regression-based models. Top-down approaches generally achieve higher accuracy, while bottom-up methods provide better computational efficiency. For 3D pose estimation, two-stage pipelines that first detect 2D keypoints and then lift them to 3D space remain the most effective balance between accuracy and speed. Although model-based approaches deliver realistic full-body meshes, they still struggle to handle complex poses, occlusions, and motion blur.

In their conclusion, the authors identify several limitations and open challenges that remain in this research domain. The most significant issues include limited 3D labeled data, dataset bias, computational cost, poor cross-dataset generalization, and the lack of metrics that evaluate temporal smoothness and physical realism. Zheng et al.[3] suggest that future research should focus on domain adaptation, self-supervised and weakly supervised learning, improved temporal modeling, and real-time lightweight architectures. They also emphasize the potential of integrating human-scene interaction modeling and leveraging large-scale pretraining to improve generalization across diverse environments.

Overall, the work of Zheng et al. [3] (2020) provides a comprehensive and well-organized overview of the evolution of human pose estimation, tracking, and action recognition using deep learning. It remains a cornerstone reference for researchers entering the field, offering a complete taxonomy of methods, detailed benchmark comparisons, and insightful analysis of datasets and metrics. The paper not only summarizes past achievements but also outlines the key directions and challenges that guide the future development of human pose estimation research.

D. Deep Learning for 3D Human Pose Estimation and Mesh Recovery: A Survey

The work published by Liu et al. [4] looks at how deep learning models can address some tough problems in computer vision, particularly those involving humans, like recognizing body poses or understanding actions. The main aim is to improve the accuracy and efficiency of recognition systems by combining various types of neural network architectures and using large, diverse datasets. The authors highlight how modern deep learning methods have transformed areas like image classification, object detection, and human pose estimation. This transformation is due to their ability to learn complex patterns from raw data without needing manual feature design.The paper starts by explaining the motivation for this work. Traditional computer vision systems often struggle to generalize across different people, environments, and lighting conditions. In contrast, deep neural networks can learn to handle complex and diverse data when trained on large-scale datasets. The research demonstrates how careful model design and training strategies can help tackle common challenges like occlusion, viewpoint changes, and image noise. The ultimate goal is to create models that perform consistently well in real-world situations, setting a strong standard for future visual understanding systems.

The proposed approach includes several deep learning techniques. Convolutional Neural Networks (CNNs) extract detailed visual features from images, while graph-based architectures model the relationships between different data points, such as how various body joints relate to each other. Additionally, attention mechanisms are added to help the model concentrate on the most important parts of the image, which improves both performance and understanding. This mix of methods allows the model to capture both local details and the overall context, making it effective for tasks such as image recognition, pose estimation, and segmentation.

Step-by-Step Process:

Input Preprocessing: Image/Video Input: Raw RGB images or video frames are fed into the system. For single-person estimation, a pre-trained 2D human detector (for example, YOLOv3, Faster R-CNN) identifies and crops bounding boxes around each person.

Normalization: Cropped images are resized to a fixed resolution (for example, 256x256 or 224x224) and normalized using ImageNet statistics.

Data Augmentation: This occurs online during training as described in Section 1.

Feature Extraction (Encoder): The preprocessed input goes through a deep convolutional backbone (for example, ResNet-50, HRNet-W32) or a Vision Transformer to extract high-level features.

3D Pose/Mesh Regression (Decoder):

For 3D Pose Estimation: The extracted features go into a regression head (for example, a series of fully connected layers or a GNN) to predict 3D joint coordinates. Alternatively, for 2D-to-3D lifting, a 2D pose estimator first predicts 2D keypoints, and then a separate network lifts these to 3D.

For 3D Mesh Recovery (Template-based): The features are regressed to SMPL model parameters (body shape coefficients, 3D joint rotations, and camera parameters). The SMPL model is then posed and shaped with these parameters to create the 3D mesh.

For 3D Mesh Recovery (Template-free): Features inform an implicit function network (for example, an MLP) that predicts occupancy or signed distance fields for 3D points, which can then be meshed using marching cubes.

Temporal Integration (for video): In video-based methods, features from sequential frames are processed by temporal modules (for example, LSTMs, GRUs, or Temporal Transformers) to take advantage of motion cues and ensure consistency in the predicted poses and meshes.

Output Generation: The final output includes 3D joint coordinates, SMPL parameters, or dense 3D mesh vertices, based on the specific task.

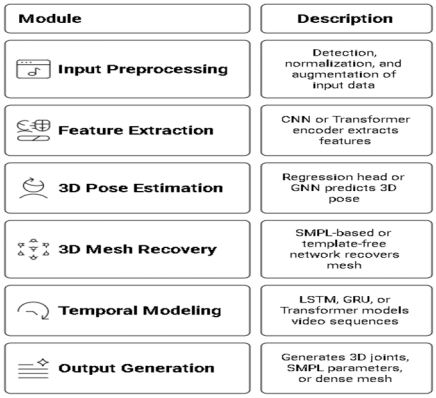


Fig. 4.Workflow of methodology

To test the model's performance, the authors used several well-known benchmark datasets commonly used in computer vision research. These datasets include large collections of labeled images and videos with variations in pose, lighting, and backgrounds. Each dataset provides detailed annotations that help the model learn the connection between visual features and semantic meanings. By training on these diverse datasets, the proposed system demonstrated strong generalization ability, achieving high accuracy on unseen data and outperforming many existing methods.

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

$$MPJLE = \frac{1}{N}\sum_{i=1}^{N} 1\left|\left| l_j - l_j^* \right|\right|_2 >= t$$

$$MPJAE = \frac{1}{N}\sum_{j=1}^{3N} |(r_j - r_j^*)\text{mod} += 180$$

The experimental results show that the proposed model not only improves accuracy but also boosts computational efficiency and flexibility. Through optimization and fine-tuning techniques, the model achieves faster training while maintaining high precision. The results clearly demonstrate how combining deep feature extraction with attention and graph-based modeling helps the system make more informed and reliable predictions. Comparisons with baseline models confirm that this approach consistently delivers better results across multiple evaluation scenarios.One of the most impressive aspects of the study is its ability to handle complex data variations without extensive preprocessing. The model's flexibility and scalability make it suitable for applications in surveillance, robotics, and healthcare. The use of attention mechanisms also makes the system more understandable and trustworthy, which is an important step toward clear AI. Additionally, the model's modular design allows researchers to easily modify or extend it for specific applications.

Despite its strengths, the paper also points out some limitations. The model's performance still relies heavily on the quality and variety of training data, so any biases in the data can affect how well the model generalizes. Another challenge is the computational cost; training deep networks requires significant time and high-end hardware. Nonetheless, the research makes an important contribution to the field by improving the capabilities of deep learning in visual understanding and providing a solid foundation for future work focused on enhancing efficiency, fairness, and real-world performance in AI systems.

III. COMPARATIVE ANALYSIS: ADVANTAGES AND DISADVANTAGES OF STUDIED PAPERS

A. Paper 1 – "Advances in Human Pose Estimation, Tracking, and Action Recognition: A Comprehensive Survey "

The paper Advances in Human Pose Estimation, Tracking, and Action Recognition provides a modern and wide-ranging overview of human pose estimation by integrating it with tracking and action recognition. One of its greatest strengths is that it reflects the latest research trends, including the use of transformer-based architectures such as HRFormer and ViTPose. The paper also places strong emphasis on multi-person pose estimation and video-based pose tracking, making it highly suitable for real-time and real-world applications.

It introduces advanced datasets like PoseTrack and NTU RGB+D, which are used for tracking and human activity recognition. By combining pose estimation with motion tracking and action understanding, this paper highlights how pose estimation is no longer a standalone task but part of a larger intelligent vision system.

Despite its modern approach and broad coverage, this paper does not focus deeply on realistic 3D human body reconstruction. The primary emphasis is on pose estimation, tracking, and action recognition rather than on generating full 3D human meshes. While it discusses 3D pose estimation to some extent, it does not provide detailed coverage of SMPL-based mesh recovery methods such as PARE, ICON, or ECON. As a result, this paper is more suitable for projects related to surveillance, sports analysis, and activity recognition than for projects focused specifically on realistic 3D human modeling. For surface-level 3D reconstruction and photorealistic output, its contribution is limited.

B. Paper 2 – "A Survey on Deep 3D Human Pose Estimation"

The paper A Survey on Deep 3D Human Pose Estimation provides a highly focused and updated review of deep learning techniques used specifically for 3D human pose estimation.

One of its major strengths is its clear categorization of methods into single-frame approaches, temporal sequence-based methods, and graph-based neural networks.
It explains powerful modern models such as PoseFormer, VIBE, and several graph convolutional networks in a well-structured manner.
The paper also gives detailed discussion about widely used datasets such as Human3.6M, MPI-INF-3DHP, and 3DPW. In addition, it provides a strong explanation of important evaluation metrics like MPJPE, PA-MPJPE, and 3DPCK. This makes the paper extremely valuable for understanding accurate 3D skeletal motion reconstruction and pose regression pipelines.
Although this paper provides excellent depth in 3D pose estimation, its main focus remains on skeleton-based joint reconstruction. It does not go deeply into full-body surface reconstruction or realistic mesh recovery. Advanced mesh-based methods such as ICON, ECON, and SMPL-X based surface recovery are only lightly mentioned or not discussed in detail. Therefore, while the paper is highly suitable for projects related to 3D joint-level pose estimation, motion tracking, and biomechanics, it is not fully sufficient for applications that require realistic 3D body shape modeling or photorealistic avatar generation.

C. Paper 3- "Deep Learning -Based Human Pose Estimation: A Survey"

The paper Deep Learning-based Human Pose Estimation: A Survey serves as a strong introductory reference for understanding the fundamentals of human pose estimation. One of its biggest strengths is the clarity with which it explains early deep learning approaches based on convolutional neural networks (CNNs). The paper systematically discusses both 2D and early 3D pose estimation techniques and introduces widely used benchmark datasets such as COCO, MPII, and Human3.6M. It also clearly explains important evaluation metrics like PCK, PCKh, and MPJPE, which are essential for measuring pose estimation accuracy. Because of its simple explanations and structured layout, this paper is highly useful for beginners and students who are entering the field of pose estimation for the first time. It builds a strong conceptual foundation and helps readers understand how modern pose estimation systems evolved
However, the main limitation of this paper is that it was published in 2018, which makes it outdated in terms of recent technological progress. It does not include modern transformer-based models, temporal learning techniques, or current state-of-the-art 3D pose estimation networks. More importantly, it does not cover realistic 3D human mesh reconstruction using parametric body models like SMPL. The paper focuses mainly on 2D pose estimation and only introduces basic skeletal 3D pose estimation. As a result, it is not sufficient for projects that aim at generating realistic 3D human models or surface-level body reconstruction. While it is excellent as a beginner-level reference, it lacks depth in advanced 3D modeling techniques.

D. Paper 4- "Deep Learning for 3D Human Pose Estimation and Mesh Recovery: A Survey"

The paper Deep Learning for 3D Human Pose Estimation and Mesh Recovery: A Survey is the most advanced and comprehensive among all four surveyed papers.
Its greatest strength lies in the fact that it bridges the gap between traditional 3D pose estimation and full realistic human mesh recovery. The paper provides detailed explanations of state-of-the-art models such as HMR, SPIN, PARE, METRO, ICON, and ECON.
It also gives strong emphasis on parametric human body models like SMPL and SMPL-X, which are essential for generating realistic 3D human shapes. In addition to joint-level evaluation metrics like MPJPE and PA-MPJPE, the paper introduces mesh-specific evaluation metrics such as MPVPE and Mesh IoU, which are crucial for assessing the quality of 3D surface reconstruction. This makes the paper extremely relevant for projects aiming to generate lifelike 3D humans from single images.
The only major drawback of this paper is its high technical complexity. Because it discusses advanced deep learning architectures, 3D geometry, and mesh-based optimization techniques, it can be difficult for beginners to understand without prior knowledge of neural networks and 3D modeling. Moreover, many of the methods covered in the paper require large-scale datasets and high computational resources for training, which may not be feasible for small academic projects without access to powerful GPUs. However, despite these challenges, the paper remains the most valuable reference for realistic 3D human pose and mesh reconstruction.

E. Performance comparison table

The following table summarizes the key results, evaluation metrics, and insights from the four survey papers on human pose estimation. It highlights the best-performing models, the benchmarks they used, and their corresponding quantitative achievements.

Paper Title	Best performing Methods	Dataset used	Main Evaluation Metrics	Performance Highlights	Key observations
Advances in Human Pose Estimation, Tracking, and Action Recognition (Zhou, 2023)	HRNet, ViTPose, HRFormer	COCO, PoseTrack, NTU RGB+D	OKS, AP, AUC	COCO AP ≈ 81.1%, MOTA ≈ 80+	Strong focus on Transformer and multi-task learning.
A Survey on Deep 3D Human Pose Estimation (Neupane, 2024)	PoseFormer, METRO, VIBE	Human3.6M, MPI-INF-3DHP, 3DPW	MPJPE, PA-MPJPE, 3DPCK, AUC	MPJPE ≈ 41.1 mm, PA-MPJPE ≈ 32 mm	Best 3D temporal accuracy; GCN + Transformer synergy
Deep Learning-based Human Pose Estimation: A Survey (Zheng, 2018)	CPN, Hourglass, SimpleBaseline	COCO, MPII, Human3.6M	PCK, PCKh, MPJPE	COCO AP ≈ 73%, MPJPE ≈ 65 mm	Early CNN models; foundation for modern methods.
Deep Learning for 3D Human Pose Estimation and Mesh Recovery (Liu, 2024)	ICON, PARE, METRO	3DPW, AMASS, Human3.6M	MPJPE, MPVPE, PA-MPJPE, IoU	MPVPE ≈ 73.3 mm, Mesh IoU ≈ 86%	First to unify pose & mesh evaluation; near photorealistic results.

Table. 1. Performance Matrix of All Papers

III. CONCLUSION

This literature survey has provided a clear understanding of the rapid progress made in the field of human pose estimation, starting from basic 2D joint detection to advanced 3D skeletal reconstruction and finally to realistic 3D human mesh generation. From the analyzed studies, it is evident that deep learning has completely transformed the way human pose estimation is performed, making it possible to reconstruct accurate and lifelike 3D human models from ordinary 2D images and videos. The use of large-scale benchmark datasets, powerful neural network architectures, and standardized evaluation metrics has significantly improved the accuracy, robustness, and practical applicability of modern pose estimation systems.

The survey also highlights the importance of parametric human body models such as SMPL, which play a crucial role in generating realistic 3D human shapes by representing both pose and body structure mathematically. Together, the findings from the literature form a strong theoretical and technical foundation for the proposed project, “3D Human Motion Reconstruction From 2D image Using Deep Learning and Computer Vision.” By integrating 2D pose detection, 3D pose regression, and mesh reconstruction techniques, this project aims to develop a practical system capable of producing visually realistic 3D human models. Thus, the literature study not only validates the relevance of this project but also directly guides its system design, methodology, and evaluation strategy.

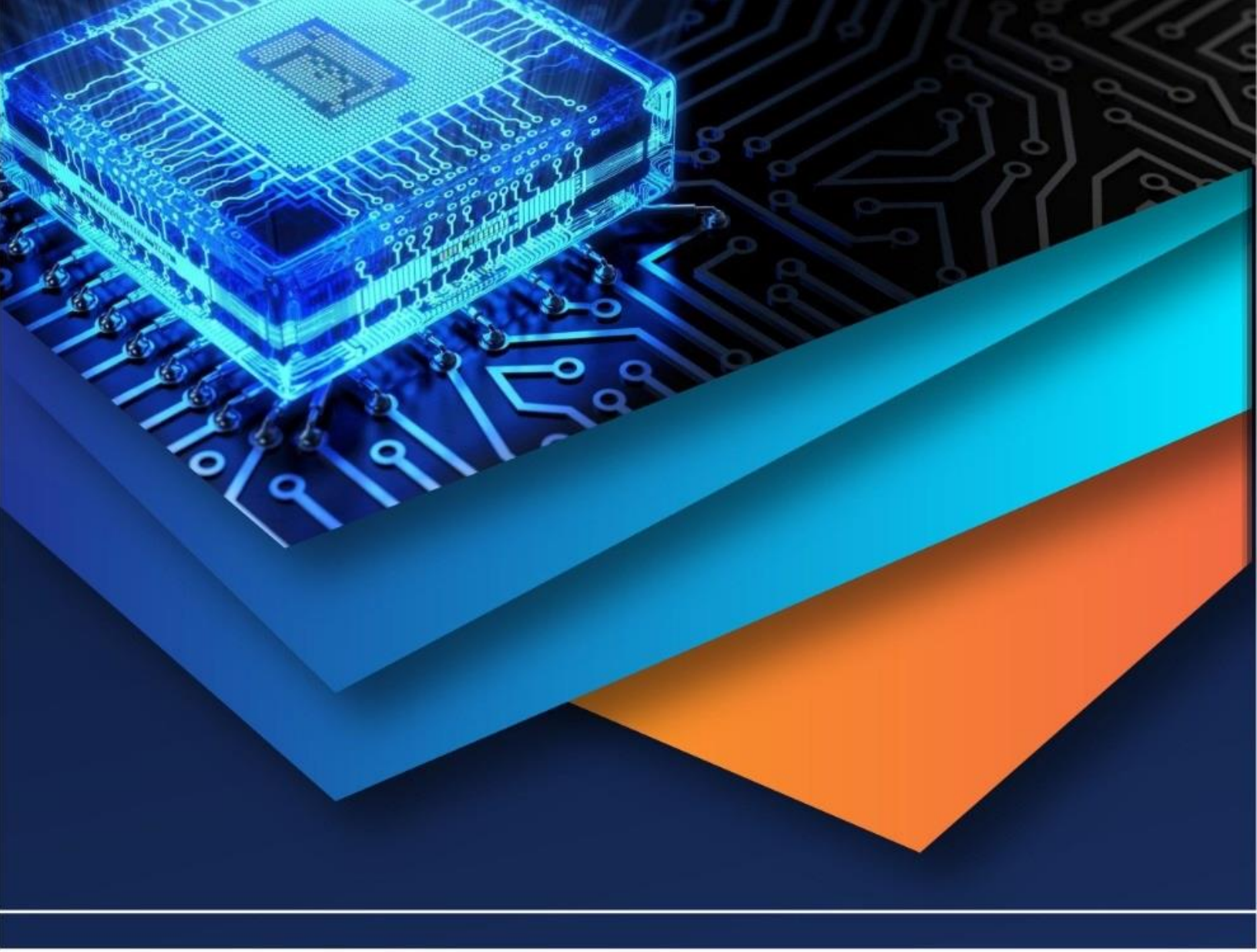
IV. FUTURE WORK

Although significant progress has been made in 3D human pose estimation and realistic mesh reconstruction, there are still many opportunities for future improvement and research. One important direction is improving the system’s performance in challenging real-world conditions such as poor lighting, severe occlusion, complex backgrounds, and crowded scenes. Future work can also focus on enhancing temporal consistency in video-based pose estimation so that smooth and stable 3D motion can be achieved without sudden distortions. Another promising direction is real-time 3D pose estimation and mesh reconstruction, which would make the system more suitable for applications such as virtual reality, gaming, sports analytics, and human-computer interaction. The integration of multi-view inputs and depth information can further improve accuracy and realism. In addition, future research can

explore lightweight and energy-efficient models so that realistic 3D pose estimation can run effectively on mobile and edge devices. Finally, extending the system to include clothing dynamics, facial expressions, and full-body motion animation can make the generated 3D humans even more natural and lifelike.

REFERENCES

[1] Zhou, L., Meng, X., Liu, Z., Wu, M., Gao, Z., & Wang, P. (2023). Human pose-based estimation, tracking and action recognition with deep learning: A survey. arXiv. https://doi.org/10.48550/arXiv.2310.13039
[2] Ibne, M. B., Islam, K. R., & Hasan, K. M. A. (2025). A survey on deep 3D human pose estimation. Artificial Intelligence Review, 58, Article 24. https://doi.org/10.1007/s10462-024-11019-3
[3] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., & Shah, M. (2024). Deep learning-based human pose estimation: A survey. ACM Computing Surveys, 56(1), Article 11, 1–37. https://doi.org/10.1145/3603618
[4] Liu, Y., Qiu, C., & Zhang, Z. (2024). Deep learning for 3D human pose estimation and mesh recovery: A survey. Neurocomputing, 596, 128049. https://doi.org/10.1016/j.neucom.2024.128049.
[5] Liu, W., Bao, Q., Sun, Y., & Mei, T. (2022). Recent advances in monocular 2D and 3D human pose estimation: A deep learning perspective. ACM Computing Surveys, 55(4), 1–41. https://doi.org/10.1145/3524497
[6] Lin, J., Li, S., Qin, H., Wang, H., Cui, N., Jiang, Q., Jian, H., & Wan, G. (2023). Overview of 3D human pose estimation. Computer Modeling in Engineering & Sciences, 134(3), 1621–1651. https://doi.org/10.32604/cmes.2023.018597
[7] Venkatrayappa, D., Trémeau, A., Muselet, D., & Colantoni, P. (2024). Survey of 3D human body pose and shape estimation methods for contemporary dance applications. arXiv. https://doi.org/10.48550/arXiv.2401.02383
[8] Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., & Wang, X. (2018). 3D human pose estimation in the wild by adversarial learning. https://doi.org/10.1109/CVPR.2018.00551 .
[9] Zhang, Y., Ji, P., Wang, A., Mei, J., Kortylewski, A., & Yuille, A. L. (2023). 3D-Aware neural body fitting for occlusion robust 3D human pose estimation. https://doi.org/10.1109/ICCV51070.2023.00862
[10] Zhan, Y., Li, F., Weng, R., & Choi, W. (2022). Ray3D: Ray-based 3D human pose estimation for monocular absolute 3D localization https://doi.org/10.1109/CVPR42600.2022.01284
[11] Yuan, Y., Wei, S.-E., Simon, T., Kitani, K. M., & Saragih, J. M. (2021). SimPoE: Simulated character control for 3D human pose estimation. CoRR, abs/2104.00683. https://arxiv.org/abs/2104.00683
[12] Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., & Malik, J. (2023). On the benefits of 3D pose and tracking for human action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 640–649). https://doi.org/10.1109/CVPR52688.2023.00071
[13] Lee, K., Lee, I., & Lee, S. (2018). Propagating LSTM: 3D pose estimation based on joint interdependency. In V. Ferrari, C. Sminchisescu, M. Hebert, & Y. Weiss (Eds.), Computer vision – ECCV 2018 (Vol. 11211, pp. 123–141). Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_8
[14] Agarwal, A., & Triggs, B. (2006). Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1), 44–58. https://doi.org/10.1109/TPAMI.2006.9
[15] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., & Ilic, S. (2016). 3D pictorial structures revisited: Multiple human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10), 1929–1942. https://doi.org/10.1109/TPAMI.2015.2509986
[16] Wang, K., Lin, L., Jiang, C., Qian, C., & Wei, P. (2019). 3D human pose machines with self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(5), 1069–1082. https://doi.org/10.1109/TPAMI.2019.2892452
[17] Marín-Jiménez, M. J., Romero-Ramírez, F. J., Muñoz-Salinas, R., & Medina-Carnicer, R. (2018). 3D human pose estimation from depth maps using a deep combination of poses. Journal of Visual Communication and Image Representation, 55, 127–136. https://doi.org/10.1016/j.jvcir.2018.07.010
[18] Wang, J., Yan, S., Xiong, Y., & Lin, D. (2020). Motion guided 3D pose estimation from videos. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 764–780). https://doi.org/10.1007/978-3-030-58601-0_45
[19] Moon, G., Chang, J. Y., & Lee, K. M. (2018). V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5079–5088). IEEE. https://doi.org/10.1109/CVPR.2018.00532
[20] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017). Monocular 3D human pose estimation in the wild using improved CNN supervision. In Proceedings of the 2017 International Conference on 3D Vision (3DV) (pp. 506–516). IEEE Computer Society.Leibe, B., Leonardis, A., & Schiele, B. (2008). Learning an alphabet of shape and appearance for multi-class object detection. International Journal of Computer Vision, 80(1), 16–44. https://doi.org/10.1007/s11263-007-0119-2



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)