TYPE Original Research
PUBLISHED 05 March 2025
DOI 10.3389/fnbot.2025.1531894

OPEN ACCESS

EDITED BY
Xin Jin,
Yunnan University, China

REVIEWED BY
Hui Liu,
University of Bremen, Germany
Szymon Wojciechowski,
Wrocław University of Science and Technology, Poland

*CORRESPONDENCE
Hui Wang
✉ 07049@gench.edu.cn

RECEIVED 21 November 2024
ACCEPTED 17 February 2025
PUBLISHED 05 March 2025

CITATION
Liu B and Wang H (2025) PoseRL-Net: human pose analysis for motion training guided by robot vision. *Front. Neurorobot.* 19:1531894. doi: 10.3389/fnbot.2025.1531894

PoseRL-Net: human pose analysis for motion training guided by robot vision

Bin Liu and Hui Wang*

Department of Physical Education, College of Education, Shanghai Jianqiao University, Shanghai, China

Objective: To address the limitations of traditional methods in human pose recognition, such as occlusions, lighting variations, and motion continuity, particularly in complex dynamic environments for seamless human-robot interaction.

Method: We propose PoseRL-Net, a deep learning-based pose recognition model that enhances accuracy and robustness in human pose estimation. PoseRL-Net integrates multiple components, including a Spatial-Temporal Graph Convolutional Network (STGCN), attention mechanism, Gated Recurrent Unit (GRU) module, pose refinement, and symmetry constraints. The STGCN extracts spatial and temporal features, the attention mechanism focuses on key pose features, the GRU ensures temporal consistency, and the refinement and symmetry constraints improve structural plausibility and stability.

Results: Extensive experiments conducted on the Human3.6M and MPI-INF-3DHP datasets demonstrate that PoseRL-Net outperforms existing state-of-the-art models on key metrics such as MPIPE and P-MPIPE, showcasing superior performance across various pose recognition tasks.

Conclusion: PoseRL-Net not only improves pose estimation accuracy but also provides crucial support for intelligent decision-making and motion planning in robots operating in dynamic and complex scenarios, offering significant practical value for collaborative robotics.

KEYWORDS

human pose estimation, 3D skeleton modeling, spatial-temporal graph convolution, attention mechanism, robot-assisted motion analysis

1 Introduction

In collaborative robotics and intelligent systems, the accuracy of human pose recognition significantly impacts the naturalness and safety of human-machine interactions, establishing it as a core technology for automation systems (Hernández et al., 2021; Liu and Wang, 2021). With the rapid advancement of deep learning and computer vision, pose recognition applications have expanded beyond robot control and monitoring to include augmented reality, sports analysis, and intelligent surveillance (Fan et al., 2022; Desmarais et al., 2021). Additionally, human pose analysis encompasses both external sensing technologies, such as vision-based systems, and internal sensing technologies, such as wearable sensor-based approaches. These two paradigms offer complementary strengths and enable a wide range of applications.

Vision-based pose recognition aims to precisely predict the position and relative relationships of various keypoints by analyzing skeletal information in videos or images (Kim et al., 2023; Liu et al., 2022). However, this task presents numerous challenges, including lighting changes (Lee and Ahn, 2020), partial occlusions (Hernández et al., 2021), diverse poses (Wang et al., 2021), and continuous motion. With the widespread adoption of deep learning, models based on Convolutional Neural Networks (CNN) (Desmarais et al., 2021), Graph Convolutional Networks (GCN) (Li et al., 2021), and Recurrent Neural Networks (RNN) (Zhang et al., 2020) have become mainstream. Methods such as OpenPose (Cao et al., 2017), AlphaPose (Fang et al., 2022), and PoseNet (Kendall et al., 2015) have shown promising results in single and multi-person settings, but they still face limitations in handling complex, dynamic scenes. Moreover, motion continuity and robustness remain challenging due to the inherent nature of visual occlusions, scene clutter, and real-time processing requirements.

Meanwhile, wearable sensor-based approaches, which use biosignals captured from devices such as smart bands, smartphones, or knee bandages, offer a privacy-preserving and unobtrusive alternative for pose analysis. These methods are particularly effective for tasks like gait parameter estimation (Hartmann et al., 2024), high-level human activity recognition (Hartmann et al., 2022), and feature space reduction for efficient classification (Hartmann et al., 2021). For example, Hartmann et al. introduced motion units for generalized sequence modeling of human activities, which address challenges such as motion continuity and robustness (Hartmann et al., 2023). Additionally, frameworks such as ASK have shown how multimodal data fusion can improve activity recognition performance, combining wearable sensor data with contextual information (Hartmann et al., 2020). Furthermore, the development of real-time wearable HAR systems demonstrates the feasibility of meeting strict latency requirements in dynamic environments (Hartmann et al., 2024). Wearable sensors excel in capturing fine-grained motion data, enabling accurate classification of activities such as standing, sitting, walking, jogging, and other locomotion patterns.

While vision-based approaches excel in capturing spatial and contextual information, sensor-based systems are highly robust to occlusions and lighting variations, making them complementary solutions for human pose analysis (Bai et al., 2024). For example, insights from wearable sensor technologies, such as motion continuity modeling and high-level feature extraction, could inspire the development of robust temporal models for vision-based systems. Similarly, multimodal systems that integrate visual and biosignal data hold promise for enhancing robustness, accuracy, and generalization across diverse scenarios. Combining these paradigms offers an opportunity to leverage the strengths of both approaches and address their respective limitations.

Human pose recognition, encompassing both external and internal sensing technologies, has broad applications in modern human-machine interaction, intelligent monitoring, and health management. In collaborative robotics, precise pose recognition enhances robots' understanding and responsiveness to human behavior, enabling them to better adapt to dynamic collaborative environments and achieve more natural interactions. In intelligent surveillance, pose recognition supports anomaly detection, crowd flow analysis, and public safety management. Additionally, in healthcare, pose recognition assists in motion assessment, posture correction, and rehabilitation training, while wearable sensors enable early detection of gait impairments or fall risks in elderly individuals (Hartmann et al., 2024). These technologies also find applications in interactive sports coaching and personalized activity tracking.

Despite significant progress, pose recognition technology still faces challenges in handling occlusions in complex scenes, enhancing motion continuity and robustness, and meeting real-time requirements. For instance, wearable sensing technologies have made strides in addressing motion continuity and real-time processing, yet their reliance on sensor placement limits generalization across datasets. Vision-based approaches, while strong in spatial feature extraction, require innovations in temporal modeling and multimodal data integration to match the robustness of sensor-based systems. Therefore, developing precise, stable, and efficient models that combine external and internal sensing modalities remains a critical research direction. Further innovations in multimodal data fusion, spatial-temporal feature extraction, and cross-domain adaptability will drive advancements in pose recognition and human activity analysis, ultimately enabling more reliable and interpretable systems for real-world applications.

PoseRL-Net is structured with key components to enhance performance: (1) A Spatial-Temporal Graph Convolutional Network (STGCN) module, which extracts spatiotemporal features from the skeletal structure by representing joints as graph nodes and skeletal connections as edges. Through graph convolution, this module learns the temporal relationships between joints, effectively capturing dynamic changes during movements, and improving pose estimation accuracy in complex, real-world scenarios, particularly in environments with occlusion or unpredictable actions. (2) An Attention Mechanism, which dynamically adjusts the weights of each joint, enabling the model to focus on the most critical joints during an action. This mechanism not only enhances recognition robustness in complex, dynamic scenes but also improves model interpretability by highlighting the most important joints in real-time, allowing the model to better handle occlusions and interactions between multiple individuals. (3) A Pose Refinement module with Symmetry Constraints, which refines predictions by enforcing symmetry between left and right joints, ensuring the physical plausibility of the predicted poses. This module reduces errors caused by asymmetric pose predictions and guarantees that the generated poses align with the natural movement patterns of the human body, enhancing both the accuracy and realism of pose estimations.

Contributions of this paper are as follows:

- This paper designs a convolutional network based on spatiotemporal graphs, which effectively extracts spatiotemporal features from the skeletal structure, addressing the challenges of dynamic, collaborative, and occluded environments, thereby enhancing recognition performance in complex real-world scenes.

- The introduction of an attention mechanism allows the model to dynamically focus on the most critical joints, improving the robustness of pose recognition in highly dynamic and interactive scenarios while offering better interpretability by emphasizing key features.
- Through the integration of a posture optimization module with symmetry constraints, the model ensures physically plausible and accurate pose predictions, significantly improving stability and generalization ability in posture recognition tasks, even in challenging environments.

2 Related work

2.1 Robot vision system structure

In collaborative environments, machine vision systems enable robots to perceive their surroundings comprehensively, assisting them in accurately recognizing poses and understanding actions in dynamic and complex scenarios (Turaga et al., 2008), which in turn allows for more precise decision-making and proactive planning (Narneg et al., 2024). Robot vision systems typically integrate various types of sensors, including monocular cameras, stereo cameras, depth cameras, and RGB-D cameras, each with its own advantages and limitations suited to different application contexts.

Monocular cameras are low-cost and easy to set up, making them suitable for basic visual functions in simple environments (Zhang et al., 2021). However, due to limitations in viewing angles and sensitivity to lighting changes, monocular cameras may struggle with robustness and accuracy in complex collaborative scenes. Stereo cameras, on the other hand, provide richer stereo vision information and have higher robustness, though their feature matching and calibration are challenging, computationally intensive, and prone to motion blur.

Depth cameras offer real-time 3D depth information and are highly resilient to lighting and shadow variations, allowing for stable detection under various lighting conditions, albeit with relatively low resolution (Dong et al., 2022; Hao et al., 2024). In comparison, RGB-D cameras (such as Kinect and TOF cameras), which combine color and depth information, achieve high-precision recognition even in complex backgrounds, making them particularly effective in collaborative scenarios with uneven lighting and multiple occlusions (Ning et al., 2024). RGB-D cameras excel in environmental adaptability and real-time performance, effectively discerning occlusion relationships between objects and identifying key features.

In complex collaborative environments, researchers have explored various sensor applications. For instance, Abdelsalam et al. (2024) employed a ZED stereo camera to capture 3D point cloud data in collaborative spaces, constructing unlabelled voxel grids and marking key elements using joint position information, achieving stable pose recognition in complex backgrounds. D'Antonio et al. (2021) utilized a Kinect camera to capture human joint information and employed a bounding circle method to correct joint displacements, enabling accurate recognition in motion. These studies highlight the potential of different vision sensors in collaborative robotic settings, demonstrating how adapting to sensor characteristics enables efficient and accurate human pose recognition in dynamic environments.

2.2 Appearance features

Appearance features primarily refer to visual attributes such as color, texture, and shape, which play essential roles in image processing. Robot vision systems analyze appearance features in images to recognize key parts of the human body (Kocabas et al., 2021), thereby enabling understanding of poses and capturing actions. However, in complex collaborative environments, recognition based on appearance features is easily affected by lighting changes, shadows, and skin color differences, which can degrade recognition performance.

Color features are a core component of appearance features and are commonly extracted through color histograms or color moments. However, variations in lighting conditions often distort color features, impacting recognition accuracy. To address this, Al Naser et al. (2022) proposed a new algorithm combining the Otsu method and the YCrCb color space, fusing thermal and color information for body part detection. Compared to the traditional OpenPose method, this approach significantly improves recognition speed and reduces the impact of lighting and skin tone variations. Additionally, Zabalza et al. (2019) developed a vision module based on low-cost cameras that uses color detection in the HSV color space, enabling robots to perceive environmental changes more accurately and detect nearby obstacles, thereby enhancing recognition precision under varying lighting and motion conditions.

In practical applications, to maintain high appearance feature recognition performance in environments with significant lighting and shadow changes, researchers typically enhance system robustness through various preprocessing and feature fusion methods. With advancements in machine vision technology, appearance feature processing techniques are continually optimized, enabling robots to capture human poses more efficiently in dynamic environments.

2.3 Local features

Local features play a crucial role in robotic vision systems, especially in complex environments with varying lighting conditions or occlusions, where local features are often more robust than appearance features (Bazzani et al., 2013). Local features describe specific details in an image, typically including edges, corners, and texture information, which help robots accurately recognize and locate key parts of the human body.

Common local feature extraction methods include Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005), and Oriented FAST and Rotated BRIEF (ORB) (Bansal et al., 2021). SIFT identifies key points in images at different scales and orientations, providing good resistance to occlusion, making it suitable for dynamic scenes. In contrast, ORB combines the FAST feature detection with the BRIEF descriptor, significantly enhancing

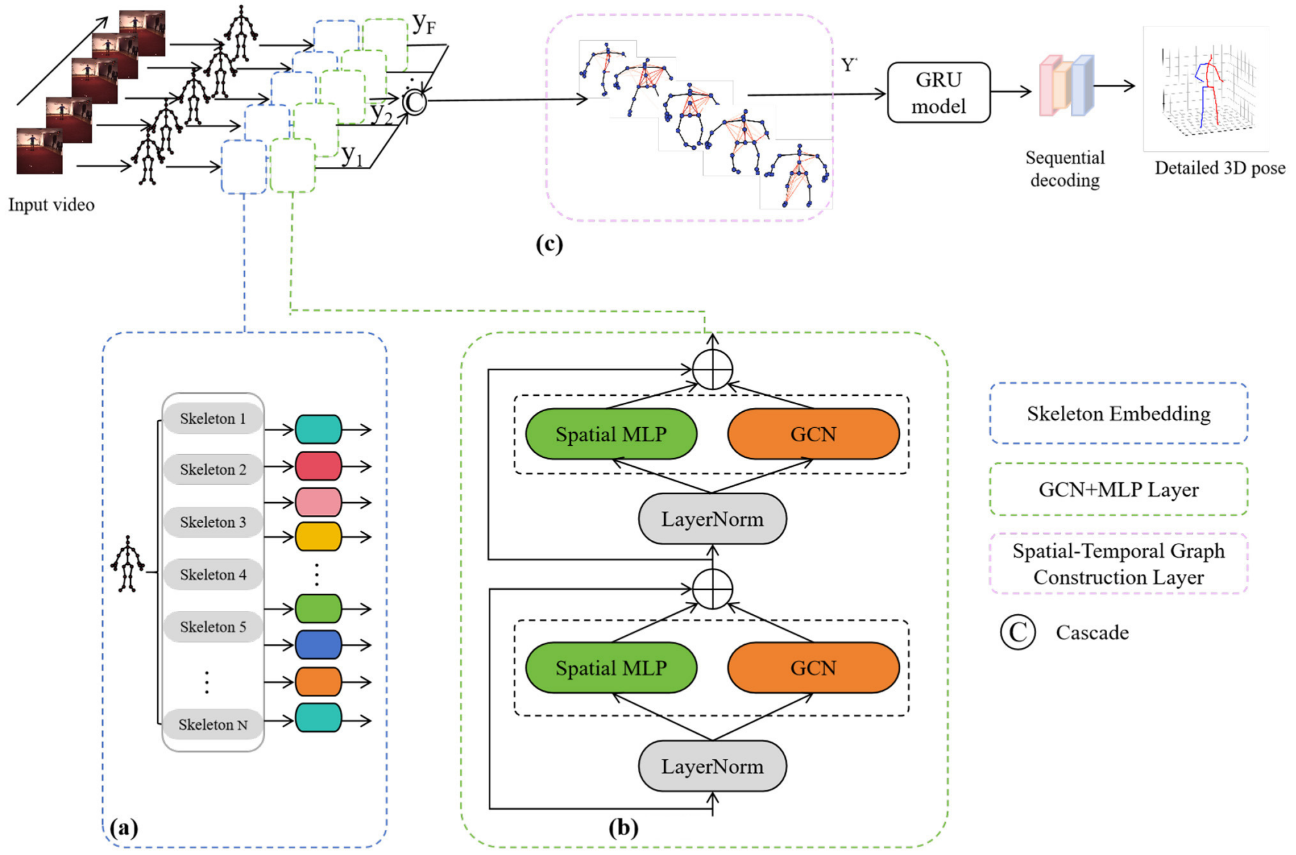

FIGURE 1
Overview of the proposed PoseRL-Net architecture. **(A)** Skeleton embedding. **(B)** GCN+MLP layer. **(C)** Spatial-temporal graph construction layer.

computational speed and making it ideal for real-time applications. In environments with significant lighting and viewpoint variations, HOG extracts feature information by calculating the gradient direction distribution across different regions of an image, offering high invariance to lighting changes.

In complex collaborative environments, researchers have further improved local feature extraction and matching methods. For instance, Vinay et al. (2015) proposed an interactive face recognition framework based on ORB, incorporating kernel principal component analysis to address nonlinear factors, significantly enhancing recognition accuracy under occlusion conditions. Wu et al. (2017) utilized HOG to extract skeleton feature matrices and proposed a Rotational and Projective Skeleton Signature (RPSS), which demonstrated good real-time performance and robustness, even when spatiotemporal information in action sequences was insufficient.

3 Method

In this paper, we propose a PoseRL-Net-based robot vision-guided method for recognizing and analyzing human posture and movement in training scenarios, with the experimental architecture shown in Figure 1. By constructing a Spatial-Temporal Graph Convolutional Network (GCN), we treat each human joint as a graph node, with edges representing spatial (skeletal) and temporal (action frame sequence) connections. This method combines the features of Multi-Layer Perceptron (MLP) and the structure of GCN, enabling efficient feature extraction and aggregation of posture data across both spatial and temporal dimensions. The model includes a posture encoding module, spatial-temporal graph convolution layer, motion prediction module, and a posture optimization step, ultimately generating guidance for robot-assisted motion training.

3.1 Spatial-temporal graph construction

In the spatial-temporal graph construction, human posture is modeled as a graph structure $G = (V, E)$. Nodes (V) represent the positions of human joints, with each node corresponding to a specific joint. Assuming the human body has M joints, and given an action sequence of T frames, the model contains $T \times M$ nodes. The edges (E) are categorized into two types, describing the spatial and temporal relationships between joints. Spatial edges connect related joints within each frame, reflecting the skeletal structure, such as the connection between the knee and hip. Temporal edges connect the same joint across consecutive frames to capture motion trajectories and temporal continuity. This structured graph allows information transfer between nodes, enabling each joint to perceive information from its neighboring joints and capture temporal action changes.

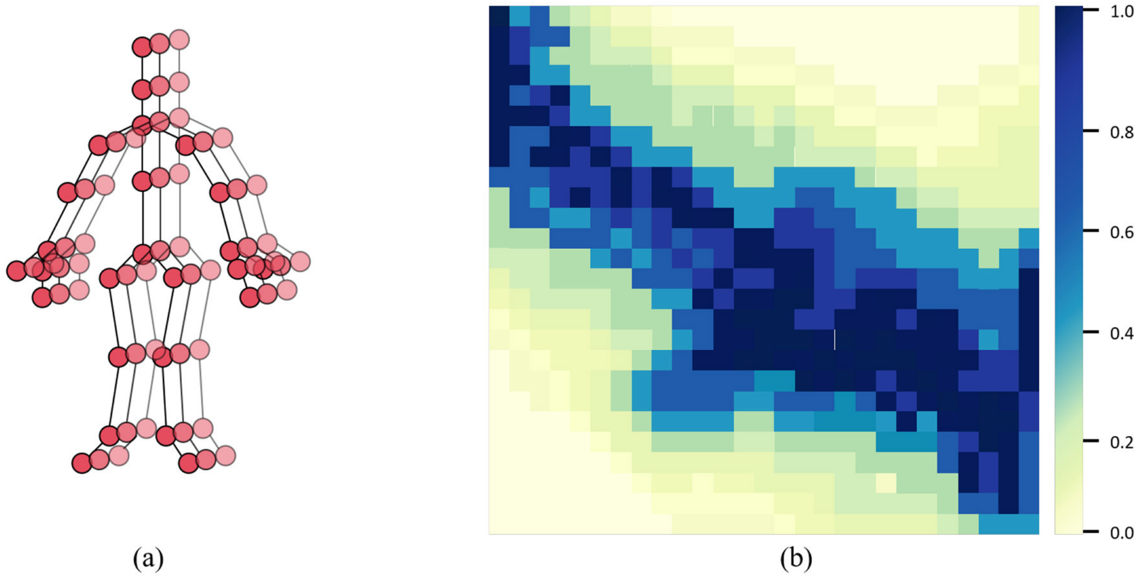

FIGURE 2
(A) Physical and symmetrical connections in the human skeletal graph. **(B)** Adjacency matrix used in the GCN module of PoseRL-Net, where different colors represent different types of bone connections.

The spatial-temporal graph construction can be expressed using the graph Laplacian operator as follows:

$$G = (V, E, W) \tag{1}$$

where $V = \{v_{ti} \mid t = 1, \dots, T; i = 1, \dots, M\}$ represents the set of all nodes, with each v_{ti} denoting the i-th joint node at frame t. $E = \{e_{ij}\}$ denotes the set of edges, defining the connections between nodes. $W = (w_{ij})_{N\times N}$ is the adjacency matrix, where $N = T \times M$ represents the total number of nodes in the graph.

The adjacency matrix W is defined as:

$$w_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{if } (i,j) \notin E \end{cases} \tag{2}$$

We split the edge set E into spatial and temporal edges. Within each frame t, the skeletal structure is viewed as a static graph, with all adjacent joint nodes connected according to the skeletal relationship. For each joint i, a temporal edge is created between the nodes v_{ti} and $v_{(t+1)i}$ across consecutive frames t and $t + 1$, representing the joint's positional change over time. To enhance the feature extraction capabilities of the graph convolution, we apply a normalized graph Laplacian matrix:

$$L = I_N - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{3}$$

where I_N is the $N \times N$ identity matrix, and D is the degree matrix with diagonal elements $D_{ii} = \sum_j w_{ij}$, representing the degree of node i. This spatial-temporal graph construction method enables the model to efficiently aggregate spatial and temporal information of joints in the graph convolution layer, providing a more accurate and comprehensive feature representation for 3D posture prediction.

3.2 Graph convolution layer

The Graph Convolution Layer (GCN) is a type of neural network layer designed to process graph-structured data by aggregating information within each node's neighborhood. This allows the central node to integrate information from its adjacent nodes, thereby enhancing the feature representation of each node. In PoseRL-Net, the joint graph structure is constructed through the spatial-temporal graph, after which the feature of each joint is enriched through the GCN layer. This process not only preserves the overall structural information of human joints but also captures temporal action patterns. During graph convolution, each node's features are propagated and merged using the adjacency and degree matrices, capturing complex spatial and temporal relationships. For human pose estimation, this operation enables information propagation between nodes, linking local joint features with the global human skeletal structure, thereby enhancing the model's expressive power and prediction accuracy.

In the GCN module of our PoseRL-Net, Figure 2A shows the joint positions of the human body, while the adjacency matrix in Figure 2B represents the skeletal connections between each pair of joints. The primary operation in the GCN layer is to aggregate and transform each node's features with those of its neighboring nodes. Given a graph $G = (V, E)$:

$$H^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}\right) \tag{4}$$

where: $H^{(l)}$: Node feature matrix at the l-th layer, with dimensions $N \times C_l$, where N is the number of nodes, and C_l is the feature dimension at layer l. $H^{(l+1)}$: Node feature matrix at the $l + 1$-th layer.$\widetilde{A} = A + I_N$: Adjacency matrix with self-connections, where A is the original adjacency matrix, and I_N is the identity matrix, representing each node's self-connection. $\widetilde{D}$: Degree matrix, where the diagonal elements are defined as $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}.W^{(l)}$: Learnable weight matrix for the l-th layer, used for linear transformation, mapping features from C_l to $C_{l+1}.\sigma$: Activation function, using ReLU as the activation function.

The GCN layer leverages graph convolution to extract spatial and temporal information between nodes. The spatial-temporal graph construction stage generates an adjacency matrix $\widetilde{A}$ that contains both spatial and temporal relationships, which the GCN layer then uses to aggregate features within each node's neighborhood. Within the same frame, spatial edges in the GCN layer connect joint nodes to learn the spatial relationships between joints, aiding the model in understanding the spatial structure of the human body. Between adjacent frames, temporal edges connect the same joint's positions across time, capturing the joint's motion trajectory and dynamic characteristics. This design helps the model learn the continuity and evolution of actions. By stacking multiple GCN layers, the model gradually aggregates spatial and temporal features at different scales, enabling a comprehensive understanding of human pose.

3.3 Hierarchical local-global feature extraction

The purpose of local-global feature extraction is to progressively aggregate local joint information into global pose features through a hierarchical approach. This process is achieved via graph pooling and graph upsampling. Specifically, the model begins with fine-grained joint-level features, gradually pooling them to form holistic information about the human body. Then, based on the global information, the model performs upsampling to recover detailed features, leading to more accurate pose predictions.

Initially, starting from the skeleton graph, we use pooling operations to aggregate information from individual joints into more abstract local features, representing different parts of the body. After several layers of pooling, the model obtains a low-resolution global feature map that captures the overall structure of the human body. Based on this global feature, the model progressively upsamples and incorporates lower-level details to generate more precise pose predictions. The features at each layer are integrated during upsampling, enabling collaborative learning of both local and global information.

Through this series of operations, the model can accurately understand the overall structure of the human body while preserving local details, thereby enhancing the accuracy of 3D pose estimation.

In the graph pooling operation at layer l, we aggregate fine-grained features into coarse-grained features. Assuming the input feature is $H^{(l)}$ and the pooled feature is $H^{(l+1)}$, the pooling process can be defined as:

$$H^{(l+1)} = P^{(l)} H^{(l)} \tag{5}$$

where $H^{(l)}$ is the node feature matrix at layer l, with dimensions $N^{(l)} \times C^{(l)}$, where $N^{(l)}$ is the number of nodes, and $C^{(l)}$ is the feature dimension at layer l. $P^{(l)}$ is the pooling matrix that controls the aggregation of features from layer l to layer $l+1$. $H^{(l+1)}$ is the feature matrix at layer $l+1$, with dimensions $N^{(l+1)} \times C^{(l+1)}$.

During the upsampling process, we progressively restore low-resolution features to higher resolutions to retain local detail information. Assuming the upsampled feature at layer l is $H^{(l)}$:

$$H^{(l-1)} = U^{(l)} H^{(l)} + H^{(l-1)} \tag{6}$$

where $H^{(l)}$ is the feature matrix at layer l, with dimensions $N^{(l)} \times C^{(l)}$. $U^{(l)}$ is the upsampling matrix that maps features from layer l to layer $l-1$. $H^{(l-1)}$ is the upsampled feature, which is fused with the pre-pooled feature at layer $l-1$ to restore details.

The hierarchical local-global feature extraction significantly enhances the model's understanding of human poses. The pooling operation allows the model to capture the overall structure of joints at a coarse-grained level, while the upsampling process retains fine-grained details, enabling the model to accurately capture subtle differences in movements during 3D pose prediction. This local-global feature extraction mechanism not only improves the model's expressive capability but also increases its adaptability and robustness to complex human actions.

3.4 Motion prediction and pose refinement

Through motion prediction, the model extends a given sequence of actions over time, capturing pose variations at different time steps. Pose refinement, on the other hand, adjusts pose parameters to ensure that the predicted results conform to natural human movement patterns, thereby enhancing both the accuracy and stability of the model.

The task of motion prediction involves temporally extending the observed sequence of poses to generate future pose information. To achieve this, the model first predicts the next frame's pose based on several past frames. This process utilizes a sequence-based deep learning network, such as a Gated Recurrent Unit (GRU), as illustrated in Figure 3. GRU enables the extension of short-term motion dynamics into long-term predictions. The core idea of motion prediction is to model the trend of joint position changes over time to forecast future joint positions:

$$\hat{\Phi}_{t+1} = f(\Phi_t, \Phi_{t-1}, \ldots, \Phi_{t-k}; \theta)$$

where $\hat{\Phi}_{t+1}$ represents the predicted pose at time $t+1$, $\Phi_t, \Phi_{t-1}, \ldots, \Phi_{t-k}$ are the historical poses from time $t-k$ to t, f is the motion prediction function, and θ denotes the model parameters. This approach allows the model to make long-term predictions of future poses based on historical data.

After motion prediction, pose refinement is applied to adjust the predicted 3D poses to align with realistic human motion and symmetry constraints. The goal of pose refinement is to optimize the prediction to ensure that the movements are physically plausible and conform to natural human structures, avoiding abrupt or anatomically inaccurate poses. This optimization typically uses specific loss functions, such as joint symmetry loss and smoothness loss, to constrain the predicted pose. Pose

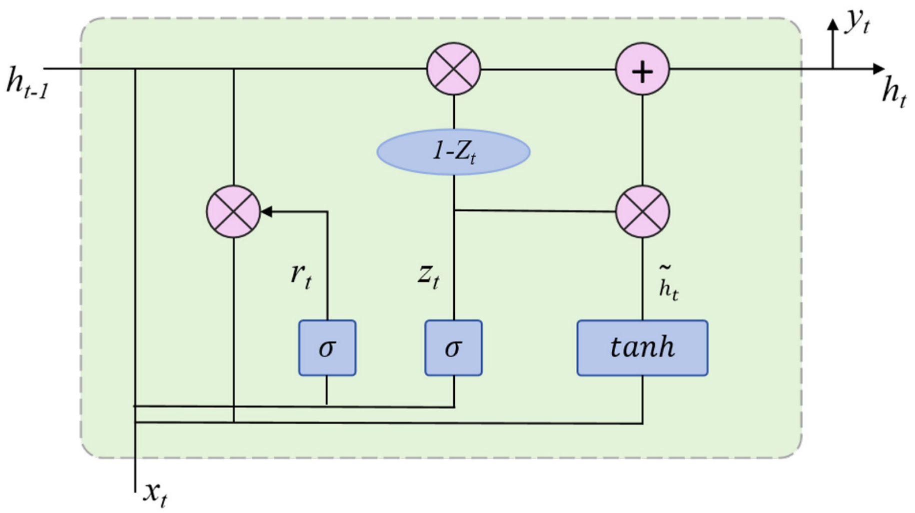

FIGURE 3
GRU module.

refinement effectively reduces the noise in predictions, thus improving the accuracy of motion forecasting.

Pose refinement primarily involves two aspects: smoothness constraint and symmetry constraint. The smoothness constraint ensures that joint movement remains consistent over time, minimizing sudden jumps, while the symmetry constraint maintains similar length and structure between symmetric joints, preserving human body symmetry.

To quantify these objectives, we design three loss functions: pose loss, smoothness loss, and symmetry loss. Each loss function targets a different aspect of model optimization, working in tandem to improve the quality of pose predictions. The specific formulas are as follows:

Pose loss L_p: This loss function measures the error between the predicted 3D joint coordinates and the ground-truth coordinates, ensuring the accuracy of the predicted pose:

$$L_p = \sum_{t=1}^{T}\sum_{i=1}^{M}\left\|\hat{\phi}_{t,i} - \phi_{t,i}\right\|^2$$

Smoothness loss L_d: This loss function constrains the change in joint positions between consecutive frames, ensuring temporal smoothness of the pose:

$$L_d = \sum_{t=2}^{T}\sum_{i=1}^{M}\left\|\hat{\phi}_{t,i} - \hat{\phi}_{t-1,i}\right\|^2$$

Symmetry loss L_s: This loss function ensures consistent relative positions and structure between symmetric joints (e.g., left and right hands, left and right legs), increasing the physical plausibility of human movement:

$$L_s = \sum_{t=1}^{T}\sum_{b\in\mathcal{B}}\left\|\left\|\hat{B}_{t,b}\right\| - \left\|\hat{B}_{t,C(b)}\right\|\right\|^2$$

where T represents the number of time steps, M is the number of joints, and $\hat{\phi}_{t,i}$ and $\phi_{t,i}$ denote the predicted and ground-truth positions of joint i at time t, respectively. In the smoothness loss, $\hat{\phi}_{t,i}$ and $\hat{\phi}_{t-1,i}$ represent the predicted positions of joint i at times t and $t-1$. In the symmetry loss, $\hat{B}_{t,b}$ and $\hat{B}_{t,C(b)}$ represent the bone vectors of symmetric joint pairs.

These loss functions work together to improve the model's capability to capture realistic and stable poses, ensuring that PoseRL-Net produces accurate and natural motion predictions across various scenarios.

4 Experiment

4.1 Datasets

The Human3.6M dataset (Ionescu et al., 2013) is one of the most commonly used 3D human pose datasets, primarily for research in 3D human pose estimation and action recognition. This dataset includes 11 different activity categories, such as walking, running, eating, and talking on the phone. These activities are performed by seven different subjects and captured from multiple camera views (a total of four cameras), providing a rich set of perspectives and fields of view. Human3.6M offers over 3.6 million annotated frames, with each frame containing 3D positions (in millimeters) of 17 joints.

The MPII Human Pose dataset (Mehta et al., 2017) is a standard 2D human pose estimation dataset containing over 25,000 images. Each image is annotated with 16 keypoints (including head, shoulders, elbows, knees, etc.). This dataset covers a wide range of everyday activities, including sports, daily tasks, and complex dance movements. Each image includes 2D joint positions for 16 joints, with poses captured from various scenes and diverse activity types.

4.2 Implementation details

In the experimental setup, we systematically trained and evaluated the model to verify the performance of PoseRL-Net. During training, we used the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.001, a momentum of 0.9, and a batch size of 32. The learning rate was reduced by a factor of 10 every 10 epochs. The training was conducted over 100 epochs, with early stopping applied to monitor the MPJPE metric on the validation set to prevent overfitting. The weights of the loss functions were set as follows: the weight for the pose loss L_p was 1, the weight for the smoothness loss L_d was 1, and the weight for the symmetry loss L_s was 0.01 to balance the contributions of each loss component. The input data was preprocessed with normalization, specifically using OpenCV and Numpy to scale-normalize the 3D joint data from the Human3.6M dataset. Model training and ablation experiments were conducted in the PyTorch framework, with TensorBoard used to log changes in loss and evaluation metrics in real time. The final evaluation metrics included MPJPE, P-MPJPE, MPJVE, parameter count, and Floating Point Operations (FLOPs) to provide a comprehensive analysis of the model's performance and computational complexity.

4.3 Evaluation metrics

In the experimental evaluation, we employed multiple metrics to comprehensively analyze the performance of PoseRL-Net. Specifically, two standard evaluation protocols were used on the Human3.6M dataset (Ionescu et al., 2013). Protocol #1 involves training the model on five subjects (S1, S5, S6, S7, S8) and testing on two unseen subjects (S9, S11) to evaluate the model's generalization to new individuals. This protocol reports the Mean Per Joint Position Error (MPJPE), which measures the average Euclidean distance between predicted and ground-truth 3D joint positions without any alignment. Protocol #2 applies rigid alignment using Procrustes analysis to remove differences in translation, rotation, and scale between the predicted and ground-truth 3D poses, using the same train-test split as Protocol #1. This protocol reports the Pose-processing MPJPE (P-MPJPE), which specifically evaluates the accuracy of pose reconstruction by focusing on the shape similarity of poses rather than their absolute position.

In addition to these metrics, the Mean Per Joint Velocity Error (MPJVE) was utilized in the MPI-INF-3DHP dataset evaluation to assess temporal smoothness by calculating the error in joint velocities between consecutive frames. Lower MPJVE values indicate better continuity and stability in motion sequences. MPJVE was not applied to the Human3.6M dataset, as its evaluation primarily focuses on pose accuracy rather than motion dynamics. Furthermore, computational efficiency was evaluated through the model's parameter count (Parameters, P) and floating-point operations (FLOPs).

4.4 Experimental results

We compare PoSER-NET with the most advanced 3D pose estimation models. The results are summarized in Table 1 and visualized in Figure 4 with the drawing of the reference (Li et al., 2023, 2024). This comparison highlights PoseRL-Net's superior predictive accuracy and stability in 3D pose estimation, as demonstrated by key metrics such as MPJPE. Table 1 details the performance of PoseRL-Net across various actions and benchmark models. PoseRL-Net consistently achieved lower MPJPE values across multiple actions, particularly in complex poses such as direction, eating, sitting, and walking. The model achieved an average MPJPE of 35.2, the lowest among all evaluated models, showcasing its exceptional ability to accurately capture human poses across diverse scenarios. The improved performance of PoseRL-Net is attributed to its critical components, including the Spatial-Temporal Graph Convolutional Network (STGC) for efficient spatio-temporal feature extraction, the attention mechanism for focusing on relevant pose features, and the symmetry constraint for maintaining structural consistency in pose predictions. These design choices collectively enabled PoseRL-Net to achieve precise and stable 3D pose estimations, as reflected in the reduced error metrics.

In Table 2 and Figure 5, PoseRL-Net outperformed other models across various actions, achieving the best average P-MPJPE of 34.9, indicating its superior accuracy and reliability in pose estimation tasks. For specific actions such as direction (Dir), eating (Eat), and sitting (Sit), PoseRL-Net demonstrated significant advantages over other models. For instance, it achieved P-MPJPE values of 32.5 for direction and 32.6 for eating, highlighting its ability to generalize across diverse poses and movements.

To evaluate the contribution of each component in PoseRL-Net, a series of ablation studies were conducted, systematically removing or replacing specific modules (e.g., STGC, GRU, pose refinement). These experiments, summarized in Table 3, aimed to quantify the impact of each module on the model's performance. The ablation results underscore the importance of every component, including STGC, attention mechanism, GRU, pose refinement, and symmetry constraint, in enhancing the accuracy of 3D pose estimation. MPJPE was used as the primary evaluation metric, with lower MPJPE values indicating better model accuracy. Additionally, model complexity was analyzed through parameters (in millions) and floating-point operations (FLOPs, in millions), offering insights into the computational cost associated with each variation.

To ensure the reliability of the ablation study results, a statistical evaluation was conducted. For each ablation experiment, the model was trained and tested 5 times with different random seeds, and the mean and standard deviation of MPJPE were reported. A paired t-test was performed to determine the statistical significance of the differences between the complete PoseRL-Net model and each ablated variant.

The complete PoseRL-Net model, incorporating all modules, achieved the lowest MPJPE of 48.6 (mean ± std: 48.6 ± 0.3) with 8.91M parameters and 894M FLOPs. This result validates the effectiveness of combining all components for accurate 3D pose estimation. Removing individual modules resulted in notable increases in MPJPE, emphasizing their significance:

Model 1 excluded the STGC module, resulting in an MPJPE of 52.1 ± 0.4 ($p < 0.01$ compared to the full model), highlighting STGC's critical role in capturing spatio-temporal features.

TABLE 1 Quantitative comparison of Protocol #1 (MPJPE) on the Human3.6M dataset.

Protocol #1	Dir	Dise	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	SMoke	Wait	WalkD	Walk	WalkT	Average
STGC-GNNs (He et al., 2023)	47.2	48.3	53.2	54.7	64.5	72.3	53.6	58.4	71.2	91.1	63.5	55.4	61.5	47.8	52.8	63.4
HiSSTGNN (Ma et al., 2023)	44.6	46.8	43.4	46.3	48.5	54.3	44.5	44.6	57.4	65.8	47.5	44.1	49.1	**32.4**	33.5	46.7
STGCN (Yu et al., 2017)	**33.5**	43.5	44.2	44.1	45.7	58.3	43.6	43.8	52.8	61.4	48.5	43.6	47.8	35.6	38.4	44.8
GAT (Veličković et al., 2017)	46.3	47.2	45.2	42.1	45.6	48.3	46.3	45.2	53.6	62.3	47.3	43.6	44.5	32.6	33.5	44.2
CNN-LSTM (Mutegeki and Han, 2020)	43.5	45.2	45.7	48.5	45.3	47.2	54.6	42.6	51.7	58.4	48.6	43.5	48.6	35.4	38.2	41.9
ResNet 18 (He et al., 2016)	44.2	41.5	48.6	47.5	45.7	48.3	54.2	45.2	52.1	57.5	48.2	42.3	48.7	35.5	38.3	44.5
ResNet 50 (He et al., 2016)	41.2	44.5	47.3	48.2	43.5	47.3	52.5	47.6	51.6	53.6	48.1	44.5	48.4	35.6	38.8	44.6
FCN (Martinez et al., 2017)	48.3	43.3	45.2	48.5	44.7	42.9	48.6	46.8	52.1	52.5	47.3	42.6	48.3	33.5	34.1	41.5
TCN (Pavllo et al., 2019)	47.2	44.1	42.6	47.2	45.1	47.2	47.3	48.6	53.3	54.7	45.6	42.5	42.6	38.6	41.5	46.3
SemGCN (Zhao et al., 2019)	45.5	48.3	45.8	42.6	44.6	42.8	48.9	45.9	57.6	51.6	42.6	47.3	49.1	34.6	48.6	47.6
GraphSH (Xu and Takano, 2021)	47.1	42.6	46.3	43.6	48.9	48.2	41.6	48.2	48.7	47.6	42.6	42.6	47.8	38.6	42.6	37.5
MGCN (Zou and Tang, 2021)	41.3	41.6	41.2	**37.4**	**41.5**	48.2	38.6	42.6	**46.5**	**42.8**	**34.6**	41.8	42.6	38.4	41.8	44.5
RS-Net (Hassan and Hamza, 2023)	38.5	**39.6**	42.6	41.6	42.5	**41.6**	**35.8**	**39.4**	50.4	54.2	47.1	**38.6**	48.2	36.4	39.5	41.7
MSS-Former (Zhao et al., 2024)	40.6	41.1	42.7	42.6	48.2	50.1	40.6	44.5	51.5	58.2	36.8	48.2	44.1	33.1	31.2	41.4
PoseRL-Net(Ours)	35.7	41.6	**40.5**	41.7	42.6	51.3	40.5	41.6	52.4	56.5	42.3	41.1	**42.1**	32.5	**30.6**	**35.2**

The best is in **bold**, the second is underlined.

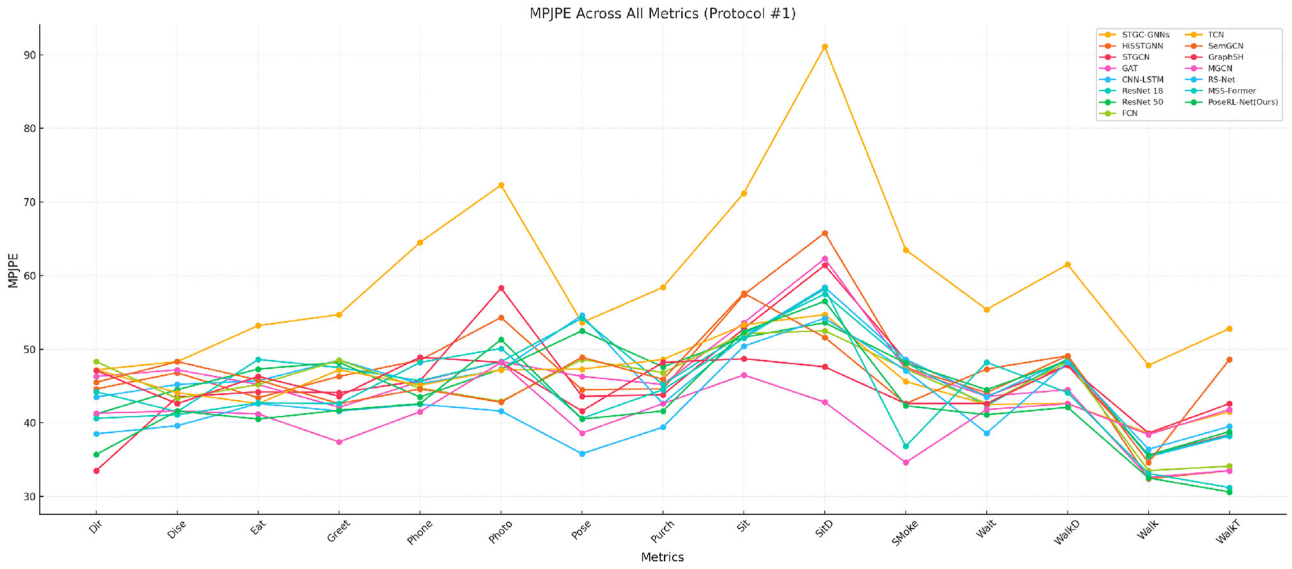

FIGURE 4
Protocol #1: MPJPE comparison across all models and metrics.

Model 2 removed the Attention Mechanism, leading to an MPJPE of 61.6 ± 0.5 ($p < 0.01$), indicating the importance of focusing on relevant pose features for accuracy. The substantial difference in MPJPE demonstrates that the attention mechanism has the most significant impact on the model's performance.

Model 3 excluded the GRU module, yielding an MPJPE of 53.6 ± 0.3 ($p < 0.05$), revealing the GRU's importance in modeling temporal dynamics for sequential pose prediction.

Model 4 omitted the pose refinement module, resulting in an MPJPE of 52.8 ± 0.4 ($p < 0.05$), demonstrating the module's contribution to noise reduction and structural consistency.

Model 5 excluded symmetry constraints, achieving an MPJPE of 51.4 ± 0.3 ($p < 0.05$), indicating its role in maintaining natural body configurations.

The statistical evaluation underscores that each module in PoseRL-Net contributes uniquely to improving 3D pose estimation, with the Attention Mechanism having the most significant effect on performance. The full model's superior performance validates the integration of STGC, attention mechanism, GRU, pose refinement, and symmetry constraints for achieving state-of-the-art results.

Table 4 presents a comparison of PoseRL-Net with state-of-the-art models on the MPI-INF-3DHP dataset, focusing on key metrics for pose estimation accuracy and efficiency. PoseRL-Net demonstrated significant improvements in accuracy compared to existing models. Specifically, it achieved the lowest MPJPE of 22.4, substantially outperforming the second-best model, MSS-Former, which had an MPJPE of 35.5. This result highlights PoseRL-Net's exceptional ability to precisely localize 3D joints. Additionally, PoseRL-Net achieved the best P-MPJPE score of 33.2 and the lowest MPJVE of 2.43, showcasing its effectiveness in maintaining alignment accuracy and smooth joint motion predictions. These results validate PoseRL-Net's ability to capture fine-grained pose details while ensuring temporal consistency in motion sequences.

In contrast, models such as Temporal Convolutional Network (TCN) and SemGCN, which lack advanced spatio-temporal and attention mechanisms, achieved higher MPJPE values of 67.2 and 70.1, respectively, demonstrating their limitations in accurately capturing complex 3D poses. GraphSH and MGCN leveraged graph-based structures and performed better with MPJPE values of 52.2 and 48.5, but they still lagged behind PoseRL-Net, highlighting its superior generalization across poses. Overall, PoseRL-Net demonstrated outstanding accuracy and efficiency on the MPI-INF-3DHP dataset, outperforming existing methods on all evaluation metrics. Its low MPJPE, P-MPJPE, and MPJVE scores underscore its robustness and applicability to 3D pose estimation in challenging scenarios. The model's balance between complexity and accuracy makes it a promising choice for real-world applications requiring precise and reliable human pose estimation.

Figure 4 provides a qualitative analysis of PoseRL-Net's performance across various human motion poses, showcasing the model's ability to accurately recognize and track human joints in 3D space. The leftmost column displays the original video frames with PoseRL-Net predictions overlaid, while the remaining columns present 3D joint estimations for different body parts from multiple perspectives. Each row represents a specific motion type, such as direction, walking, stooping, and appel. In the visualized 3D plots, the predicted poses (blue) are compared with the ground truth poses (red), allowing for observation of joint alignment and accuracy. Zoomed-in areas highlight critical joints, demonstrating the model's precision in capturing fine details and maintaining smooth tracking across frames. Overall, Figure 6 effectively illustrates PoseRL-Net's robustness and high accuracy in handling complex human poses, making it suitable for applications in motion analysis and human pose estimation.

TABLE 2 Quantitative comparison of Protocol #2 (P-MPJPE) on the Human3.6M dataset.

Protocol #2	Dir	Dise	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	SMoke	Wait	WalkD	Walk	WalkT	Avg.
STGC-GNNs (He et al., 2023)	34.1	35.6	42.6	48.6	36.4	**39.5**	41.4	42.3	47.2	52.5	48.2	36.8	45.2	39.5	37.3	41.8
HiSSTGNN (Ma et al., 2023)	33.5	41.5	46.4	37.2	45.2	42.6	44.2	33.5	42.3	55.8	37.1	33.0	47.2	38.0	47.2	41.6
STGCN (Yu et al., 2017)	31.0	34.8	38.1	35.6	43.6	43.9	43.6	42.5	41.6	51.2	47.5	36.2	49.5	34.7	47.6	41.4
GAT (Veličković et al., 2017)	36.5	34.5	37.5	37.2	42.6	45.8	42.3	40.6	43.1	57.5	49.2	38.4	47.2	**25.6**	35.6	40.9
CNN-LSTM (Mutegeki and Han, 2020)	35.5	36.1	36.2	**34.4**	45.7	56.0	34.4	42.5	45.0	69.4	42.6	33.8	39.1	36.5	43.1	42.0
ResNet 18 (He et al., 2016)	42.7	43.2	64.4	35.2	51.0	51.4	47.3	33.6	46.8	49.0	47.2	38.2	37.8	37.1	41.5	44.4
ResNet 50 (He et al., 2016)	47.5	42.6	54.2	38.2	47.2	52.6	**31.6**	32.1	42.6	52.6	37.4	34.2	**35.8**	26.9	31.9	40.5
FCN (Martinez et al., 2017)	41.2	42.6	34.7	35.5	41.6	42.2	38.7	33.6	56.5	35.4	38.2	45.0	42.6	37.6	36.5	40.1
TCN (Pavllo et al., 2019)	39.5	44.2	42.6	47.0	45.7	47.6	48.2	45.3	49.6	41.5	35.4	41.5	43.5	38.4	38.2	43.2
SemGCN (Zhao et al., 2019)	38.4	41.2	36.1	47.2	36.2	42.7	41.6	45.2	44.6	51.8	39.7	42.6	44.7	37.5	36.4	41.7
GraphSH (Xu and Takano, 2021)	41.5	42.6	37.5	48.3	44.5	41.7	42.5	36.5	**41.5**	47.5	34.1	45.1	47.2	29.4	34.5	41.0
MGCN (Zou and Tang, 2021)	47.2	41.6	35.1	42.8	47.2	45.6	41.2	45.2	44.1	**41.2**	42.5	41.2	42.4	35.5	34.5	41.8
RS-Net (Hassan and Hamza, 2023)	35.4	45.2	36.1	41.1	44.1	42.5	41.6	44.5	42.5	41.4	42.4	47.5	44.1	31.5	**31.5**	40.8
MSS-Former (Zhao et al., 2024)	33.6	48.5	39.5	42.5	45.6	42.8	45.6	41.2	47.2	42.6	45.7	43.3	41.5	36.2	34.5	42.0
PoseRL-Net(Ours)	**32.5**	**32.5**	**32.6**	36.2	**33.1**	42.3	36.7	**31.4**	42.1	41.3	**33.5**	**32.1**	36.5	28.4	31.6	**34.9**

The best is in **bold**, the second is underlined.

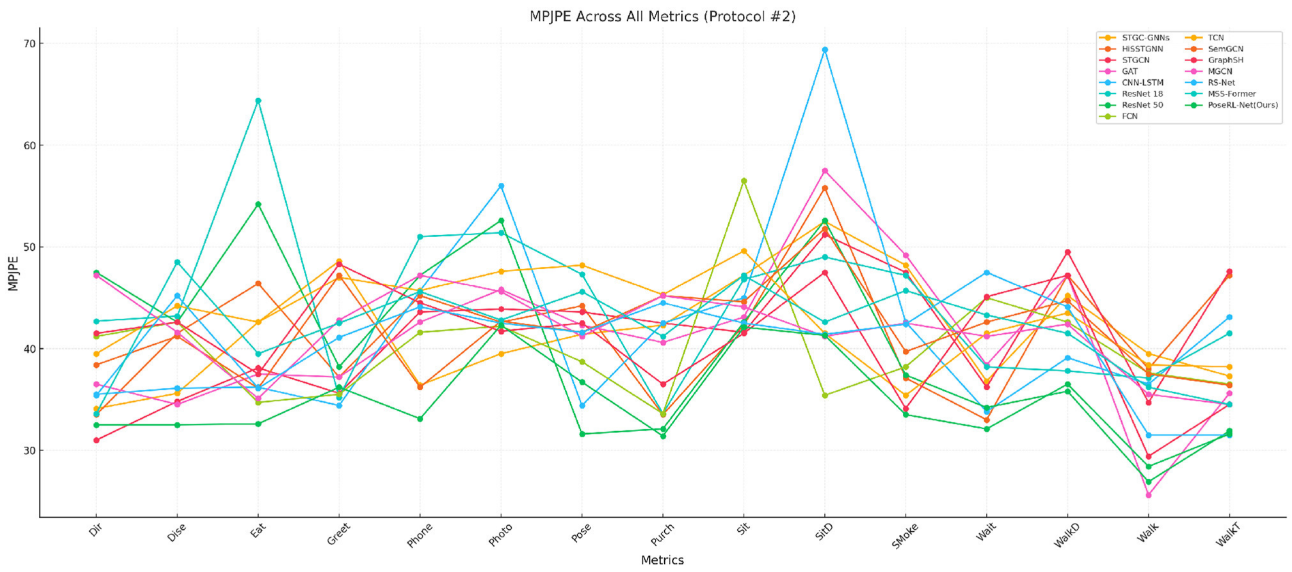

FIGURE 5
Protocol #2: MPJPE comparison across all models and metrics.

TABLE 3 Ablation analysis of each module.

Model	STGC	Attention mechanism	GRU	Pose refinement	Symmetry constraint	Parameters (M)	FLOPs (M)	MPJPE
PoseRL-Net	✓	✓	✓	✓	✓	8.91	894	48.6
Model 1		✓	✓	✓	✓	6.02	472	52.1
Model 2	✓		✓	✓	✓	7.62	847	61.6
Model 3	✓	✓		✓	✓	8.42	465	53.6
Model 4	✓	✓	✓		✓	8.69	731	52.8
Model 5	✓	✓	✓	✓		8.82	684	51.4

TABLE 4 Accuracy of different models on the MPI-INF-3DHP dataset.

Method	Parameters (M)	FLOPs (M)	MPJPE	P-MPJPE	MPJVE
TCN (Pavllo et al., 2019)	88.5	78.3	67.2	39.4	2.96
SemGCN (Zhao et al., 2019)	89.6	80.4	70.1	36.2	2.87
GraphSH (Xu and Takano, 2021)	90.4	81.5	52.2	35.6	2.76
MGCN (Zou and Tang, 2021)	90.1	81.2	48.5	34.4	2.54
RS-Net (Hassan and Hamza, 2023)	90.5	82.5	62.7	33.6	2.64
MSS-Former (Zhao et al., 2024)	90.4	81.1	35.5	34.1	2.56
PoseRL-Net(Ours)	91.5	81.4	22.4	33.2	2.43

5 Discussion and conclusion

This paper proposes a robot posture recognition method based on machine vision, PoseRL-Net, which aims to improve the 3D posture recognition performance of robots in complex collaborative scenarios. By combining key components such as spatiotemporal graph convolutional network (STGC), attention mechanism, GRU module, posture optimization and symmetry constraint, PoseRL-Net can effectively extract and fuse spatiotemporal features to achieve accurate recognition of human posture. We have conducted extensive experimental verification of PoseRL-Net on datasets such as Human3.6M and MPI-INF-3DHP, and compared it with existing advanced methods. The experimental results show that PoseRL-Net has achieved the best performance in multiple indicators, especially in key indicators such as MPJPE and P-MPJPE, which is significantly better than other models, demonstrating its excellent performance in posture estimation tasks.

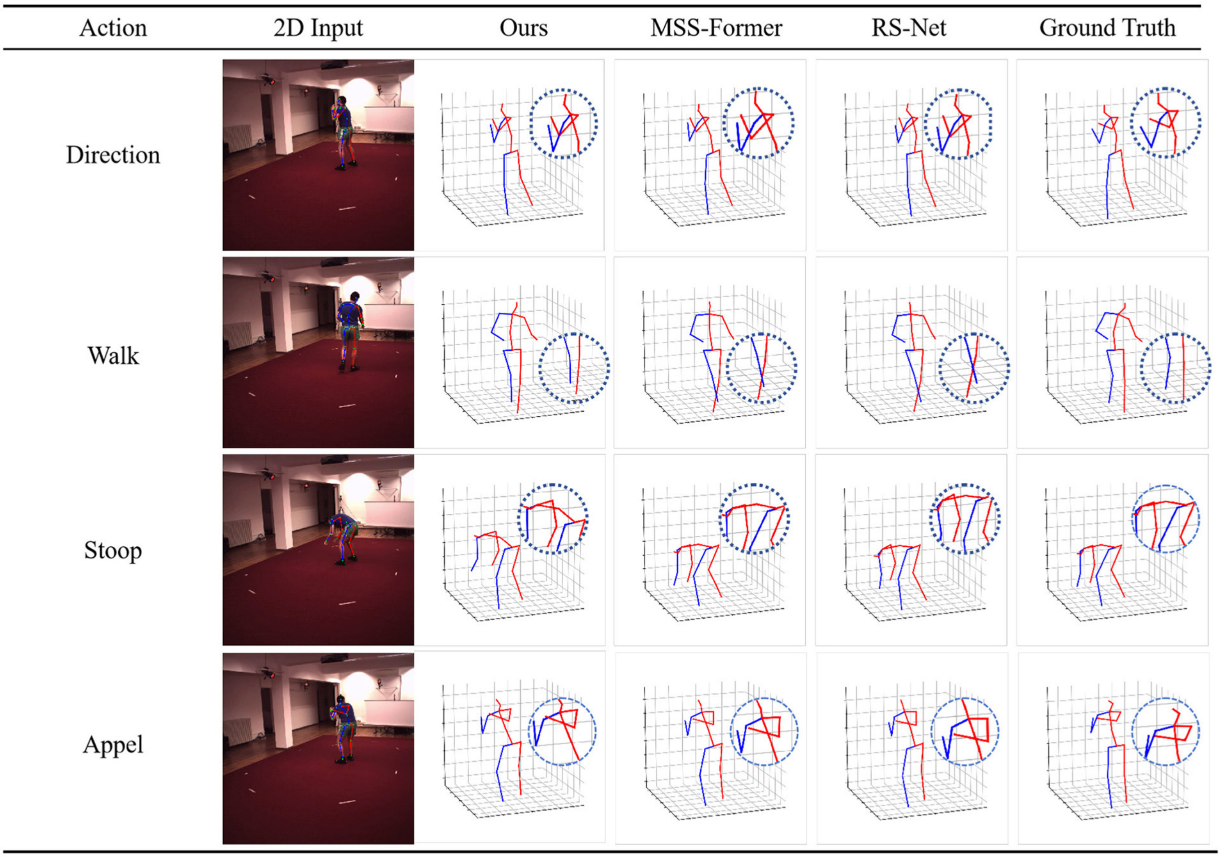

FIGURE 6
Examples of MSS-Former, RS-Net, and our 3D pose estimation model. The last column is the 3D pose of the ground truth.

The effectiveness of each module is further verified by ablation experiments, which proves the importance of modules such as spatiotemporal feature extraction, posture optimization and symmetry constraint to improve model accuracy. In addition, we also analyzed the training and inference efficiency of PoseRL-Net under different hardware conditions, proving its potential in real-time applications. In practical applications, PoseRL-Net can adapt to a variety of postures and motion types, has strong robustness and generalization capabilities, and provides reliable technical support for the intelligent decision-making and action planning of robots in complex human-machine collaborative environments.

Future research can further explore the lightweight optimization of models in low computing resource environments to achieve a wider range of application scenarios. In short, this study provides an effective solution for posture recognition under the guidance of robot vision, and lays a solid foundation for the application of collaborative robots in the field of human-machine collaboration.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

BL: Validation, Writing – review & editing. HW: Conceptualization, Data curation, Methodology, Resources, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdelsalam, A., Mansour, M., Porras, J., and Happonen, A. (2024). Depth accuracy analysis of the ZED 2I stereo camera in an indoor environment. *Rob. Auton. Syst.* 179:104753. doi: 10.1016/j.robot.2024.104753

Al Naser, I., Dahmen, J., Bdiwi, M., and Ihlenfeldt, S. (2022). "Fusion of depth, color, and thermal images towards digital twins and safe human interaction with a robot in an industrial environment," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (IEEE), 532–537. doi: 10.1109/RO-MAN53752.2022.9900548

Bai, X., Zhang, P., Yu, X., Zheng, J., Hancock, E. R., Zhou, J., et al. (2024). Learning from human attention for attribute-assisted visual recognition. *IEEE Trans. Pattern. Anal. Mach. Intell.* 46, 11152–11167. doi: 10.1109/TPAMI.2024.3458921

Bansal, M., Kumar, M., and Kumar, M. (2021). 2D object recognition: a comparative analysis of sift, surf and orb feature descriptors. *Multimed. Tools Appl.* 80, 18839–18857. doi: 10.1007/s11042-021-10646-0

Bazzani, L., Cristani, M., and Murino, V. (2013). Symmetry-driven accumulation of local features for human characterization and re-identification. *Comput. Vis. Image Understand.* 117, 130–144. doi: 10.1016/j.cviu.2012.10.008

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299. doi: 10.1109/CVPR.2017.143

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (IEEE), 886–893. doi: 10.1109/CVPR.2005.177

D'Antonio, E., Taborri, J., Mileti, I., Rossi, S., and Patané, F. (2021). Validation of a 3D markerless system for gait analysis based on openPose and two RGB webcams. *IEEE Sens. J.* 21, 17064–17075. doi: 10.1109/JSEN.2021.3081188

Desmarais, Y., Mottet, D., Slangen, P., and Montesinos, P. (2021). A review of 3D human pose estimation algorithms for markerless motion capture. *Comput. Vis. Image Underst.* 212:103275. doi: 10.1016/j.cviu.2021.103275

Dong, X., Garratt, M. A., Anavatti, S. G., and Abbass, H. A. (2022). Towards real-time monocular depth estimation for robotics: a survey. *IEEE Trans. Intell. Transp. Syst.* 23, 16940–16961. doi: 10.1109/TITS.2022.3160741

Fan, J., Zheng, P., and Li, S. (2022). Vision-based holistic scene understanding towards proactive human-robot collaboration. *Robot. Comput. Integr. Manuf.* 75:102304. doi: 10.1016/j.rcim.2021.102304

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., et al. (2022). Alphapose: whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 7157–7173. doi: 10.1109/TPAMI.2022.3222784

Hao, M., Zhang, Z., Li, L., Dong, K., Cheng, L., Tiwari, P., et al. (2024). Coarse to fine-based image–point cloud fusion network for 3D object detection. *Inf. Fusion* 112:102551. doi: 10.1016/j.inffus.2024.102551

Hartmann, Y., Klöckner, J., Deichsel, L., Paul, R. E., and Schultz, T. (2024). "Gait parameter estimation from a single privacy preserving depth sensor," in *BIOSTEC*, 637–645. doi: 10.5220/0012383700003657

Hartmann, Y., Liu, H., Lahrberg, S., and Schultz, T. (2022). "Interpretable high-level features for human activity recognition," in *Biosignals*, 40–49. doi: 10.5220/0010840500003123

Hartmann, Y., Liu, H., and Schultz, T. (2020). "Feature space reduction for multimodal human activity recognition," in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - BIOSIGNALS* (INSTICC, SciTePress), 135–140. doi: 10.5220/0008851401350140

Hartmann, Y., Liu, H., and Schultz, T. (2021). "Feature space reduction for human activity recognition based on multi-channel biosignals," in *BIOSIGNALS*, 215–222. doi: 10.5220/0010260800002865

Hartmann, Y., Liu, H., and Schultz, T. (2023). "High-level features for human activity recognition and modeling," in *Biomedical Engineering Systems and Technologies*, eds. A. C. A. Roque, D. Gracanin, R. Lorenz, A. Tsanas, N. Bier, A. Fred, and H. Gamboa (Cham: Springer Nature Switzerland), 141–163.

Hassan, M. T., and Hamza, A. B. (2023). Regular splitting graph network for 3D human pose estimation. *IEEE Trans. Image Proc.* 32, 4212–4222. doi: 10.1109/TIP.2023.3275914

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. doi: 10.1109/CVPR.2016.90

He, S., Luo, Q., Du, R., Zhao, L., He, G., Fu, H., et al. (2023). STGC-GNNS: a GNN-based traffic prediction framework with a spatial-temporal granger causality graph. *Phys. A* 623:128913. doi: 10.1016/j.physa.2023.128913

Hernández, Ó. G., Morell, V., Ramon, J. L., and Jara, C. A. (2021). Human pose detection for robotic-assisted and rehabilitation environments. *Appl. Sci.* 11:4183. doi: 10.3390/app11094183

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2013). Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 1325–1339. doi: 10.1109/TPAMI.2013.248

Kendall, A., Grimes, M., and Cipolla, R. (2015). "Posenet: a convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2938–2946. doi: 10.1109/ICCV.2015.336

Kim, J.-W., Choi, J.-Y., Ha, E.-J., and Choi, J.-H. (2023). Human pose estimation using mediapipe pose and optimization method based on a humanoid model. *Appl. Sci.* 13:2700. doi: 10.3390/app13042700

Kocabas, M., Huang, C.-H. P., Hilliges, O., and Black, M. J. (2021). "Pare: part attention regressor for 3D human body estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11127–11137. doi: 10.1109/ICCV48922.2021.01094

Lee, J., and Ahn, B. (2020). Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors* 20:2886. doi: 10.3390/s20102886

Li, J., Li, J., Wang, C., Verbeek, F. J., Schultz, T., and Liu, H. (2023). Outlier detection using iterative adaptive mini-minimum spanning tree generation with applications on medical data. *Front. Physiol.* 14:1233341. doi: 10.3389/fphys.2023.1233341

Li, J., Li, J., Wang, C., Verbeek, F. J., Schultz, T., and Liu, H. (2024). MS2OD: outlier detection using minimum spanning tree and medoid selection. *Mach. Learn.* 5:015025. doi: 10.1088/2632-2153/ad2492

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2021). Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 3316–3333. doi: 10.1109/TPAMI.2021.3053765

Liu, H., Liu, T., Zhang, Z., Sangaiah, A. K., Yang, B., and Li, Y. (2022). ARHPE: asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction. *IEEE Trans. Ind. Inf.* 18, 7107–7117. doi: 10.1109/TII.2022.3143605

Liu, H., and Wang, L. (2021). Collision-free human-robot collaboration based on context awareness. *Robot. Comput. Integr. Manuf.* 67:101997. doi: 10.1016/j.rcim.2020.101997

Lowe, D. G. (1999). "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision* (IEEE), 1150–1157. doi: 10.1109/ICCV.1999.790410

Ma, M., Xie, P., Teng, F., Wang, B., Ji, S., Zhang, J., et al. (2023). HISTGNN: hierarchical spatio-temporal graph neural network for weather forecasting. *Inf. Sci.* 648:119580. doi: 10.1016/j.ins.2023.119580

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3D human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2640–2649. doi: 10.1109/ICCV.2017.288

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., et al. (2017). "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *2017 international conference on 3D vision (3DV)* (IEEE), 506–516. doi: 10.1109/3DV.2017.00064

Mutegeki, R., and Han, D. S. (2020). "A CNN-LSTM approach to human activity recognition," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (IEEE), 362–366. doi: 10.1109/ICAIIC48513.2020.9065078

Narneg, S., Adedoja, T., Ayyalasomayajula, M. M. T., and Chintala, S. (2024). AI-driven decision support systems in management: enhancing strategic planning and execution. *Int. J. Rec. Innov. Trends Comput. Commun.* 12, 268–275.

Ning, X., Yu, Z., Li, L., Li, W., and Tiwari, P. (2024). DILF: differentiable rendering-based multi-view image–language fusion for zero-shot 3d shape understanding. *Inf. Fusion* 102:102033. doi: 10.1016/j.inffus.2023.102033

Pavllo, D., Feichtenhofer, C., Grangier, D., and Auli, M. (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*, 7753–7762. doi: 10.1109/CVPR.2019.00794

Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: a survey. *IEEE Trans. Circ. Syst. Video Technol*. 18, 1473–1488. doi: 10.1109/TCSVT.2008.2005594

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vinay, A., Shekhar, V. S., Rituparna, J., Aggrawal, T., Murthy, K. B., and Natarajan, S. (2015). Cloud based big data analytics framework for face recognition in social networks using machine learning. *Procedia Comput. Sci*. 50, 623–630. doi: 10.1016/j.procs.2015.04.095

Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., et al. (2021). Deep 3D human pose estimation: a review. *Comput. Vis. Image Underst*. 210:103225. doi: 10.1016/j.cviu.2021.103225

Wu, H., Shao, J., Xu, X., Ji, Y., Shen, F., and Shen, H. T. (2017). Recognition and detection of two-person interactive actions using automatically selected skeleton features. *IEEE Trans. Hum. Mach. Syst*. 48, 304–310. doi: 10.1109/THMS.2017.2776211

Xu, T., and 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*, 16105–16114. doi: 10.1109/CVPR46437.2021.01584

Yu, B., Yin, H., and Zhu, Z. (2017). Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Zabalza, J., Fei, Z., Wong, C., Yan, Y., Mineo, C., Yang, E., et al. (2019). Smart sensing and adaptive reasoning for enabling industrial robots with interactive human-robot capabilities in dynamic environments–a case study. *Sensors* 19:1354. doi: 10.3390/s19061354

Zhang, J., Liu, H., Chang, Q., Wang, L., and Gao, R. X. (2020). Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. *CIRP Ann*. 69, 9–12. doi: 10.1016/j.cirp.2020.04.077

Zhang, Y., Zhao, J., and Han, H. (2021). A 3D machine vision-enabled intelligent robot architecture. *Mobile Inf. Syst*. 2021:6617286. doi: 10.1155/2021/6617286

Zhao, L., Peng, X., Tian, Y., Kapadia, M., and Metaxas, D. N. (2019). "Semantic graph convolutional networks for 3D human pose regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3425–3435. doi: 10.1109/CVPR.2019.00354

Zhao, Q., Fan, X., Chen, M., Xiao, Y., Wang, X., Yeung, E. H. K., et al. (2024). MSS-former: multi-scale skeletal transformer for intelligent fall risk prediction in older adults. *IEEE Internet Things J*. 11, 33040–33052. doi: 10.1109/JIOT.2024.3420789

Zou, Z., and Tang, W. (2021). "Modulated graph convolutional network for 3D human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11477–11487. doi: 10.1109/ICCV48922.2021.01128