

Spatiotemporal Fusion Algorithm Based on ST-GCN and Attention BiLSTM: A Method Study for Diverse Motion Posture Recognition

Yonghong Guo^a, Zulakbal Abd Karim^{1b}, Yuan Lei^c

^a*Lishui University, Xueyuan Road, Liandu District, Lishui, 323060, Zhejiang, China*

^b*Universiti Pendidikan Sultan Idris, Tanjong Malim, 35900, Perak, Malaysia*

^c*KiwiBridge Ascend Education Limited, Auckland City, 0626, New Zealand*

Abstract

Posture recognition technology has broad application prospects in sports training, health monitoring, human-computer interaction, and other fields. However, it still faces challenges in accuracy and robustness in complex and changing real environments. This research proposes a multi-scenario adaptive posture recognition algorithm MAPR, which processes human posture data through spatial-temporal graph convolutional networks and attention-enhanced bidirectional long short-term memory networks, achieving effective expression of spatial structural features and temporal dynamic features. The algorithm was comprehensively evaluated on the MPII human pose dataset, comparing the performance of the complete MAPR algorithm, variants without ST-GCN, variants without Attention-BiLSTM, and the HRNet baseline model in normal and dim environments. Experimental results show that the MAPR algorithm performs excellently in normal environments, with PCKh@0.5 reaching 0.892 and AUC value reaching 0.919. More significantly, in dim environments with 40% illumination, the MAPR algorithm still maintains high performance (AUC=0.863), with a performance decrease of only 6.1%, clearly outperforming other comparison models. By analyzing the recognition difficulty of different types of actions, the research reveals the complementary value of spatial features and temporal features in posture recognition. The MAPR algorithm provides new ideas for posture recognition in diverse motion scenarios, showing significant robustness especially under environmental challenges such as changes in lighting conditions, laying a technical foundation for applications in motion analysis, sports training, and other fields.

Keywords: Spatiotemporal Graph Convolutional Networks (ST-GCN), Attention-enhanced Bidirectional Long Short-Term Memory Networks (Attention-BiLSTM), Multi-scene Adaptive Pose Recognition (MAPR), Low-light Human Pose Estimation, Sports Motion Analysis

1. Introduction

1.1. Research Background

Posture recognition technology plays an increasingly important role in modern sports science, with applications spanning professional sports training, competitive performance analysis, and sports medicine rehabilitation [1]. With the rapid development of computer vision and deep learning technologies, the demand for automated recognition and analysis of human movement posture has become particularly urgent [2]. Currently, sports training and competitive analysis still mainly rely on coaches' subjective experience and simple video playback technology, lacking precise, objective quantitative analysis tools [3]. This situation not only limits the precise optimization of athletes' technical movements but also hinders the in-depth development of sports science research toward refinement and data-driven directions.

In practical application scenarios, posture recognition faces many challenges, especially the generalization ability problem in diverse movement environments. Different sports have their own unique movement characteristics and evaluation standards, from swimming's streamlined movements to gymnastics' complex flips, from tennis's explosive swings to marathon's sustained gait patterns. These differences greatly increase the difficulty of building a universal recognition system. Traditional single-model methods usually perform well in specific movement types but have poor adaptability across movement scenarios, making it difficult to meet the comprehensive analysis needs of modern sports [4, 5, 6]. Additionally, complex environmental factors such as lighting changes, occlusion problems, and perspective changes further increase the technical threshold for accurate recognition, leading to unstable performance of existing systems in practical applications [7].

In response to the above problems, developing a universal posture recognition system that can adapt to various sports scenarios and has high-precision recognition capabilities becomes particularly important. Such a system not only needs to accurately capture the spatial relationships of human joints but also needs to understand the temporal changes of key actions and the

importance differences of specific joints in different sports events. By integrating the advantages of spatial structure analysis and temporal dynamic modeling, a comprehensive intelligent system for understanding sports postures can be constructed. Such a system will provide coaches and athletes with objective and precise technical movement analysis, assist in developing personalized training plans, optimize sports performance, and reduce injury risks. At the same time, in the field of sports science research, high-precision posture recognition will also provide reliable data foundation for cutting-edge research such as sports biomechanics and movement pattern analysis, promoting the development of sports science toward precision and personalization.

1.2. Related Work

Human posture recognition research has made significant progress in the past decade, gradually evolving from traditional visual methods to a method system dominated by deep learning [8]. Early research mainly relied on hand-designed features and geometric models. With the rise of deep learning technology, end-to-end posture recognition methods have made breakthrough progress in accuracy and real-time performance [9]. Research in this field can be divided into two main directions: 2D pose estimation and 3D pose estimation. 2D pose estimation focuses on locating human key points from single-perspective images, while 3D pose estimation further infers the spatial coordinates and rotation information of human joints [10]. In recent years, monocular camera-based posture recognition technology has received widespread attention. Researchers have proposed various methods based on convolutional neural networks, recurrent neural networks, and graph neural networks to improve the accuracy and robustness of pose estimation [11]. These technologies are not only applied to human posture recognition but also extended to specific fields such as head pose estimation [12], vocal cord pose estimation [13], and engineering machinery pose estimation [14], showing the wide applicability and strong potential of deep learning in posture recognition problems. However, in practical applications, posture recognition systems still face challenges such as insufficient data, depth ambiguity, and occlusion, especially in complex and changing sports scenarios, where recognition accuracy and generalization ability need to be improved [8].

Spatiotemporal modeling methods play a key role in the field of posture recognition, mainly reflected in robustness issues across different perspectives and scenarios. The challenge of pose estimation lies in the processing

of multi-perspective and complex postures, especially when extreme appearance changes distort discriminative features, traditional posture recognition methods produce undesirable results under cross-covariant conditions [15]. Existing research mainly develops along two technical routes: appearance-based methods and skeleton-based methods. Appearance-based methods can utilize visual context information but perform poorly in modeling large action temporal dynamics through optical flow estimation; while skeleton-based methods ignore important visual background information such as scenes and objects [16]. In response, researchers have proposed various fusion strategies, such as Optical Flow Pose Image (OFPI) combined with skeleton data temporal aggregation [17], and contrastive learning-based deep methods [18]. However, existing methods still have limitations when dealing with challenges such as large-angle head poses [19, 20], posture changes in facial expression recognition [21], and specific domains like yoga movement recognition [22] and cross-modal supervision such as through-wall radar human posture recognition [23]. These limitations are mainly manifested in three aspects: first, a single model has difficulty in handling both spatial structure and temporal changes well; second, there is a lack of effective mechanisms for key action recognition and multi-rhythm movement adaptation; third, there is insufficient generalization ability between different movement types. The cascade integration architecture proposed in this paper is precisely aimed at these challenges. Through the complementary advantages of ST-GCN and Attention-BiLSTM, it achieves comprehensive modeling of spatial relationships and temporal features while solving the adaptability problem of different sports scenarios.

To comprehensively understand the current research status in the field of posture recognition, Table 1 summarizes the representative work in this field in recent years. The table compares the characteristics of different studies in terms of application scenarios, research content, and possible shortcomings. From this, it can be seen that although existing methods have achieved good results in specific scenarios, there are still obvious limitations in terms of generalization ability in diverse sports scenarios, key moment recognition, and comprehensive modeling of spatiotemporal features. These limitations are precisely the core problems that the cascade integration architecture MAPS system proposed in this paper attempts to solve.

Table 1: Posture Recognition Method Comparison

Authors	Application Scenario	Research Content	Possible Limitations
Huang et al. [15]	Gait recognition	Proposed an integrated posture learning network based on appearance transfer (GaitApp), separating posture features and appearance features through multi-branch collaboration	Mainly focused on gait features, insufficient adaptability for complex movement postures and rapidly changing action sequences
Chen et al. [17]	Action recognition	Proposed Optical Flow Pose Image (OFPI) method, enhancing action representation through skeleton data temporal aggregation	Lacks recognition mechanism for key action moments, may perform poorly in multi-rhythm movements
Cui et al. [16]	Video action recognition	Designed Pose-Appearance Relation Network (PARNet), establishing association between human posture and image appearance	Complex three-stream network structure, difficult collaborative optimization between modules, high computational cost
Liu et al. [18]	3D human action posture recognition	Used deep contrastive learning and improved Transformer for 3D human action posture recognition	Strong dependency on noise filtering, suboptimal in coordination between spatial structure modeling and temporal analysis
Cubero et al. [24]	Posture-based gait recognition	Proposed hierarchical limb representation method, enhancing body posture description, applied to heat maps and DensePose	Mainly focuses on static posture features, insufficient modeling of dynamic temporal changes, lacks key action recognition mechanism
Zhang et al. [25]	3D pose estimation under fish-eye camera	Surveyed monocular 3D human pose estimation and action recognition, proposed new benchmark dataset based on fisheye camera	Only for specific camera type, lacks universal solution, insufficient adaptability for multiple movement types
Kan et al. [26]	Pose estimation in industrial parts grasping	Proposed contour feature matching-based pose estimation method, processing incomplete point cloud scenes	Focused on rigid objects, cannot be applied to complex pose estimation of non-rigid targets such as human body

Continued on next page

Table 1 continued

Authors	Application Scenario	Research Content	Possible Limitations
Wang et al. [21]	Facial expression recognition	Proposed unsupervised multi-source domain adaptation-based pose-robust personalized facial expression recognition framework	Limited to facial expressions, difficult to extend to full-body posture recognition, lacks modeling of temporal information
Zhou et al. [22]	Yoga posture recognition	Two-stage action recognition method based on lightweight pose estimation model, enhancing recognition accuracy through heat maps	Specific to yoga scenarios, limited generalization ability for posture recognition in high-speed movements and complex environments
Luan et al. [27]	Face recognition	Proposed contrastive learning symmetric siamese network framework (SSN), solving face recognition problems under posture changes	Focused on solving a single problem (long-tail data distribution), lacks comprehensive modeling of spatiotemporal features
Yang et al. [19]	Multi-view face detection	Designed HeadPose-Softmax curriculum learning loss function, dynamically adjusting the importance of difficult samples according to pose angles	Only for head posture, not applicable to full-body movement analysis, lacks temporal modeling ability

1.3. Our Contributions

- This research proposes a new posture recognition system called MAPR (Multi-scene Adaptive Pose Recognition), which effectively fuses spatial structural features and temporal dynamic information through cascading ST-GCN and Attention-BiLSTM, two complementary algorithms. Experiments show that compared with single models, MAPR achieves an AUC value of 0.919 in normal environments, 12.6 percentage points higher than the baseline model HRNet; in dim environments (40% illumination), the performance advantage further expands, with AUC value maintained at 0.863, 14.7 percentage points higher than HRNet, demonstrating the excellent performance of this architecture in diverse movement scenarios.
- This research systematically analyzes the recognition difficulty differences of various actions in normal and dim environments, revealing the respective advantage areas of ST-GCN and Attention-BiLSTM components. Experimental data shows that ST-GCN performs prominently

in processing complex spatial relationships. For instance, in complex action categories, models containing ST-GCN have 17mm lower MPJPE in dim environments than models without this component; while Attention-BiLSTM has unique advantages in capturing temporal information, especially when processing fast actions, models with this component show only a 6.9% decrease in KARF metric performance, compared to a 10.4% decrease for models without this component.

- This research explores the influence mechanism of lighting conditions on posture recognition performance. By analyzing the performance degradation trends of the model on various evaluation metrics, the research finds that the MAPR model shows the minimum performance degradation on all metrics (PCKh@0.5 decreased by 5.6%, KARF decreased by 6.9%, AUC decreased by 6.1%), verifying the robustness of this algorithm in countering environmental interference, providing strong support for posture recognition applications in complex real-world scenarios.

2. Methodology

2.1. Problem Statement

The core task of a posture recognition system is to accurately extract and represent human movement posture from input visual data. Human posture can be represented as a skeleton model with spatial structure, defined as a set of N key points:

$$P = \{p_i | p_i \in \mathbb{R}^3, i = 1, 2, \dots, N\} \quad (1)$$

where p_i represents the three-dimensional coordinates of the i -th joint, and N is the total number of joints in the human model (in the MPII dataset, $N = 16$). The coordinate value of each joint p_i contains three components (x_i, y_i, z_i) , representing its position in three-dimensional space. The connection relationship between joints in the skeleton structure can be represented by an adjacency matrix A :

$$A \in \mathbb{R}^{N \times N}, A_{ij} = \begin{cases} 1, & \text{if joint } i \text{ and } j \text{ are directly connected} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The element A_{ij} in matrix A indicates whether there is a direct connection between the i -th and j -th joints, and this matrix describes the topological structure of the human skeleton. In the temporal dimension, motion posture can be represented as a posture sequence:

$$S = \{P_t | t = 1, 2, \dots, T\} \quad (3)$$

where P_t is the posture representation at time step t , and T is the total length of the sequence. This sequence captures the continuous changes in posture during the motion process. For different sports such as swimming, gymnastics, or throwing events, the sequence length T may vary significantly. For different types of movements, the importance of joints varies, and importance weights can be introduced:

$$W = \{w_i | w_i \in [0, 1], i = 1, 2, \dots, N\} \quad (4)$$

where w_i represents the importance weight of the i -th joint in a specific type of movement, with values ranging from $[0, 1]$. For example, in throwing movements, arm joints may have higher weights (close to 1), while in running, leg joints have higher weights. These weights directly affect the assessment of movement quality. The movement quality evaluation function can be defined as the weighted distance between key posture points and the standard template:

$$Q(S, S^*) = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N w_i \cdot d(p_{i,t}, p_{i,t}^*) \quad (5)$$

where S^* represents the standard action sequence, $p_{i,t}$ and $p_{i,t}^*$ are the positions of the i -th joint at time t in the actual sequence and the standard sequence, respectively, $d(\cdot, \cdot)$ is a distance metric function, which can be Euclidean distance or cosine similarity, etc. A smaller Q value indicates higher movement quality, meaning the actual movement is closer to the standard movement. In practical application scenarios, due to factors such as camera angles, lighting conditions, and occlusions, the observed joint positions often contain noise and errors:

$$\hat{p}_i = p_i + \varepsilon_i \quad (6)$$

where $\hat{p}_i$ is the observed joint position, p_i is the true position, and ε_i represents observation noise, typically assumed to follow a zero-mean Gaussian

distribution $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, where σ_i^2 represents the observation variance. In sports venues, different lighting conditions and movement speeds may lead to changes in noise distribution.

Problem 1. *Based on the mathematical description above, the core problem this research needs to solve is: how to construct a robust posture recognition function f that can accurately reconstruct the true movement posture sequence S from noisy observation data $\hat{S}$, and can adapt to various sports types from swimming to gymnastics:*

$$f(\hat{S}, \mathcal{C}) \rightarrow S, \text{ such that } \min_{S^*} Q(S, S^*) < \epsilon \quad (7)$$

where $\mathcal{C}$ represents the context information of the movement type, determining the importance weight allocation of different joints and the evaluation criteria for specific movements, and ϵ is an acceptable error threshold. This function needs to consider both the posture spatial structure features described in equation (2) and the temporal dynamic features captured in equation (3), and maintain high recognition accuracy in diverse movement scenarios while dealing with noise interference described in equation (6).

2.2. Graph Convolution-Based Joint Spatial Relationship Modeling

2.2.1. Algorithm Comparison

- **Limitations of Traditional Methods:** Traditional posture recognition methods typically use convolutional neural networks (CNN) or recurrent neural networks (RNN) to process spatiotemporal features independently, making it difficult to effectively model the spatial relationships between joints. CNNs ignore the physical connection constraints between joints when processing skeleton data, while RNNs, although capable of modeling temporal dependencies, are insufficient at capturing spatial topological structures, resulting in limited recognition accuracy in complex postures and changing environments.
- **Advantages of ST-GCN:** The Spatial-Temporal Graph Convolutional Network (ST-GCN) effectively fuses spatial structure and temporal evolution information by representing the human skeleton as a graph structure and performing convolution operations directly on joints and skeletal edges, as shown in Figure 1. Its spatial partitioning strategy and motion-sensitive attention mechanism can adaptively highlight key

joints in different types of movements, improving the model’s generalization ability and robustness in diverse movement scenarios, particularly excelling in stability issues in complex environments.

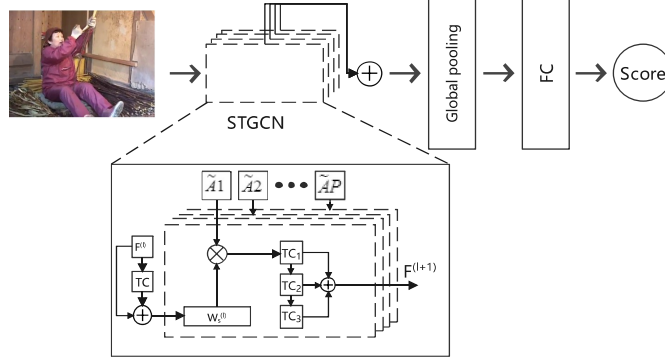


Figure 1: ST-GCN Structure

2.2.2. Spatiotemporal Graph Convolutional Network Design

For the spatial structure modeling problem in diverse movement posture recognition, the spatiotemporal graph convolutional network (ST-GCN) provides a graph structure-based solution. ST-GCN focuses on addressing the spatial relationship modeling proposed in the problem description section, particularly the skeleton structure described by equations (1) and (2). Given human posture P and its adjacency matrix A , an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be constructed, where the vertex set $\mathcal{V} = \{v_i | i = 1, 2, \dots, N\}$ corresponds to joints, and the edge set $\mathcal{E} = \{e_{ij} | (i, j) \in \mathcal{V} \times \mathcal{V}, A_{ij} = 1\}$ corresponds to skeletal connections. Based on this graph structure, ST-GCN represents the motion posture sequence as a temporal graph sequence:

$$\mathcal{G}_S = \{\mathcal{G}_t | \mathcal{G}_t = (\mathcal{V}, \mathcal{E}, F_t), t = 1, 2, \dots, T\} \quad (8)$$

where $F_t \in \mathbb{R}^{N \times C_{in}}$ is the node feature matrix at time step t , and C_{in} is the input feature dimension. For the initial input, $C_{in} = 3$, corresponding to the three-dimensional coordinates of the joints. The core of ST-GCN is the graph convolution operation, which is implemented by extending traditional convolutional neural networks to the graph domain. Based on spectral graph theory, graph convolution can be defined through the eigendecomposition

of the graph Laplacian operator. Given adjacency matrix A and diagonal degree matrix D ($D_{ii} = \sum_j A_{ij}$), the normalized graph Laplacian matrix is defined as:

$$L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U \Lambda U^T \quad (9)$$

where I_N is the $N \times N$ identity matrix, U is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues. Based on this decomposition, the graph convolution operation can be defined as:

$$\text{GConv}(F, A) = U g_\theta(\Lambda) U^T F = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}) F \quad (10)$$

where $g_\theta(\Lambda)$ is a filter in the frequency domain, which can be approximated by Chebyshev polynomials T_k , $\tilde{L} = 2L/\lambda_{\max} - I_N$ is the rescaled Laplacian matrix, θ_k are learnable parameters, and K is the order of the polynomial. To capture features in both spatial and temporal dimensions, ST-GCN introduces a spatiotemporal graph convolution module. This module first applies spatial graph convolution, followed by standard 1D convolution in the temporal dimension:

$$\begin{aligned} F^{(l+1)} &= \sigma \left(\text{BN} \left(\text{TConv} \left(\text{BN} \left(\text{SConv}(F^{(l)}, A, W_s^{(l)}) \right) \right), W_t^{(l)} \right) \right) \\ \text{SConv}(F, A, W_s) &= \sum_{p=0}^{P-1} W_s^{(p)} \cdot (D_p^{-\frac{1}{2}} A_p D_p^{-\frac{1}{2}}) \cdot F \\ \text{TConv}(F, W_t) &= F *_t W_t \end{aligned} \quad (11)$$

where $F^{(l)} \in \mathbb{R}^{N \times T \times C_l}$ is the feature tensor of the l -th layer, σ is the ReLU activation function, BN represents batch normalization, $W_s^{(l)} \in \mathbb{R}^{C_l \times C_{l+1} \times P}$ are the learnable weights for spatial convolution, $W_t^{(l)} \in \mathbb{R}^{C_{l+1} \times C_{l+1} \times K_t}$ are the learnable weights for temporal convolution, K_t is the temporal convolution kernel size, and $*_t$ represents the convolution operation in the temporal dimension. To enhance the model's ability to capture relationships between different types of joints, ST-GCN adopts a spatial partitioning strategy, subdividing the adjacency matrix A into multiple subsets:

$$A = \sum_{p=0}^{P-1} A_p, \quad \text{where } A_p \cap A_q = \emptyset, \forall p \neq q \quad (12)$$

A three-partition strategy is typically adopted ($P = 3$): A_0 for self-connection, corresponding to the joint itself; A_1 for center-outward, corresponding to connections from central joints to limb joints; A_2 for outward-center, corresponding to connections from limb joints to central joints. To capture motion patterns at different time scales, ST-GCN adopts a multi-scale temporal convolution strategy, processing different receptive fields through parallel convolution branches:

$$\begin{aligned} \text{MS-TConv}(F, \{W_t^{(k)}\}_{k=1}^K) &= \text{Concat}(\{\text{TConv}_k(F, W_t^{(k)})\}_{k=1}^K) W_{\text{fuse}} \\ \text{TConv}_k(F, W_t^{(k)}) &= F *_t W_t^{(k)} \end{aligned} \quad (13)$$

where $\{\text{TConv}_k\}_{k=1}^K$ are K temporal convolutions with different kernel sizes, and $W_{\text{fuse}} \in \mathbb{R}^{K C_{l+1} \times C_{l+1}}$ are fusion weights. To address the issue of different joint importance in different types of movements, ST-GCN introduces a motion-sensitive attention mechanism to adaptively adjust joint weights:

$$\begin{aligned} \alpha_i^{(l)} &= \text{Softmax} \left(\frac{W_q^{(l)} F_i^{(l)} \cdot (W_k^{(l)} F^{(l)})^T}{\sqrt{C_l}} \right) \\ \tilde{F}_i^{(l)} &= \alpha_i^{(l)} \cdot (W_v^{(l)} F^{(l)}) + F_i^{(l)} \end{aligned} \quad (14)$$

where $F_i^{(l)}$ is the feature of the i -th joint, $W_q^{(l)}$, $W_k^{(l)}$, and $W_v^{(l)}$ are the transformation matrices for query, key, and value, respectively, and $\alpha_i^{(l)}$ is the attention weight. The expressive power of ST-GCN can be theoretically guaranteed by the following theorem:

Theorem 1 (Expressive Power of ST-GCN). *Given a skeleton graph $\mathcal{G}$ with N joints and an ST-GCN network of sufficient depth, the network can approximate any graph signal function $f : \mathbb{R}^{N \times T \times C_{in}} \rightarrow \mathbb{R}^{N \times C_{out}}$ that satisfies spatial locality and temporal stationarity, such that:*

$$\mathbb{E}_{F \sim \mathcal{D}} [\|f(F) - \text{ST-GCN}(F)\|_2^2] < \epsilon \quad (15)$$

where $\mathcal{D}$ is the input distribution, and ϵ is an arbitrarily small positive number.

From this theorem, an important corollary about ST-GCN’s cross-movement type transfer ability can be further derived:

Corollary 1 (Movement Type Invariance). *Through spatial partitioning strategy and motion-sensitive attention mechanism, ST-GCN can learn a general representation of skeleton structure that is independent of specific movement types, while retaining movement type-specific information, achieving feature transfer ability across movement types:*

$$d_{\text{feature}}(ST\text{-}GCN(F^{(s)}), ST\text{-}GCN(F^{(t)})) < \delta \cdot d_{\text{structure}}(\mathcal{G}^{(s)}, \mathcal{G}^{(t)}) \quad (16)$$

where $F^{(s)}$ and $F^{(t)}$ are the input features of the source and target domains, respectively, $\mathcal{G}^{(s)}$ and $\mathcal{G}^{(t)}$ are the corresponding skeleton graphs, d_{feature} is the feature space distance metric, $d_{\text{structure}}$ is the graph structure distance metric, and δ is a constant related to the network parameters.

2.3. Attention Mechanism Enhanced Action Sequence Temporal Analysis

2.3.1. Algorithm Comparison

- **Limitations of Traditional Methods:** Traditional temporal modeling methods such as standard LSTM and GRU have information decay problems when processing long action sequences, making it difficult to capture long-distance dependencies. At the same time, these methods assign the same weight to all time steps, unable to distinguish between key action moments and transitional actions, resulting in poor performance in multi-rhythm movements and action continuity evaluation, with limitations particularly evident in sports analysis scenarios that require precise identification of technical action key frames.
- **Advantages of Attention-BiLSTM:** The attention mechanism enhanced bidirectional long short-term memory network (Attention-BiLSTM) effectively solves the long-distance dependency problem by processing sequence information bidirectionally, considering both past and future context, as shown in Figure 2. Its core advantage lies in the temporal attention mechanism’s ability to automatically identify and highlight key action moments, while the multi-scale temporal attention fusion design enables the model to adapt to different rhythm movement patterns, from fast explosive actions to slow sustained actions. This design makes the model excel in key action moment recognition and action continuity evaluation.

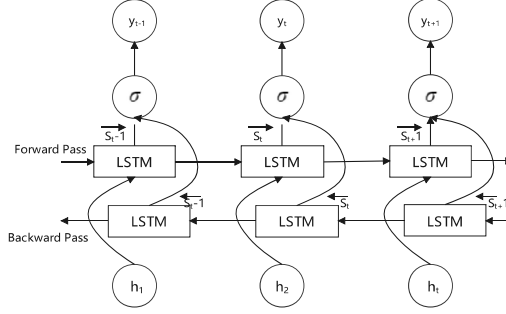


Figure 2: Attention-BiLSTM Algorithm Structure

2.3.2. Attention Enhanced Bidirectional LSTM Model

While the spatiotemporal graph convolutional network (ST-GCN) excels in spatial relationship modeling, it has limitations in capturing long-term dependencies and identifying key action moments. To address this issue, the attention mechanism enhanced bidirectional long short-term memory network (Attention-BiLSTM) is designed to process the feature sequence output by ST-GCN, focusing on temporal modeling and key moment identification. Given the feature sequence extracted by ST-GCN $H = \{h_t | h_t \in \mathbb{R}^d, t = 1, 2, \dots, T\}$, where h_t is the feature vector at time step t , and d is the feature dimension, Attention-BiLSTM first processes the entire sequence through bidirectional LSTM to capture long-term dependencies:

$$\begin{aligned}
 \vec{s}_t &= \text{LSTM}_{\text{forward}}(h_t, \vec{s}_{t-1}) \\
 \overleftarrow{s}_t &= \text{LSTM}_{\text{backward}}(h_t, \overleftarrow{s}_{t+1}) \\
 s_t &= [\vec{s}_t; \overleftarrow{s}_t]
 \end{aligned} \tag{17}$$

where $\vec{s}_t$ and $\overleftarrow{s}_t$ are the hidden states of the forward and backward LSTM at time step t , respectively, $[\cdot; \cdot]$ represents vector concatenation, $s_t \in \mathbb{R}^{2d_h}$ is the merged hidden state, and d_h is the LSTM hidden state dimension. The internal computation of the LSTM unit follows the following gating mechanism:

$$\begin{aligned}
i_t &= \sigma(W_i \cdot [s_{t-1}; h_t] + b_i) \\
f_t &= \sigma(W_f \cdot [s_{t-1}; h_t] + b_f) \\
o_t &= \sigma(W_o \cdot [s_{t-1}; h_t] + b_o) \\
\tilde{c}_t &= \tanh(W_c \cdot [s_{t-1}; h_t] + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
s_t &= o_t \odot \tanh(c_t)
\end{aligned} \tag{18}$$

where i_t , f_t , and o_t are the input gate, forget gate, and output gate, respectively, $\tilde{c}_t$ is the candidate cell state, c_t is the cell state, $\odot$ represents element-wise multiplication, σ and $\tanh$ are the sigmoid and hyperbolic tangent activation functions, respectively, and W_i , W_f , W_o , W_c and b_i , b_f , b_o , b_c are learnable parameters. To identify key action moments and assign them higher weights, Attention-BiLSTM introduces a temporal attention mechanism:

$$\begin{aligned}
e_t &= v_a^T \tanh(W_a s_t + b_a) \\
\alpha_t &= \frac{\exp(e_t)}{\sum_{j=1}^T \exp(e_j)} \\
c &= \sum_{t=1}^T \alpha_t s_t
\end{aligned} \tag{19}$$

where e_t is the attention score at time step t , α_t is the normalized attention weight, c is the context vector, and v_a , W_a , and b_a are learnable parameters. This attention mechanism can automatically identify key action moments in the movement sequence, such as jumping points, rotation moments, or racket contact points. To address the adaptation problem of multi-rhythm movements, Attention-BiLSTM introduces a multi-scale temporal attention fusion mechanism:

$$\begin{aligned}
c_k &= \sum_{t=1}^T \alpha_t^{(k)} s_t, \quad k = 1, 2, \dots, K \\
\alpha_t^{(k)} &= \frac{\exp(e_t^{(k)})}{\sum_{j=1}^T \exp(e_j^{(k)})} \\
e_t^{(k)} &= v_a^{(k)T} \tanh(W_a^{(k)} s_t + b_a^{(k)}) \\
c_{\text{multi}} &= \sum_{k=1}^K \beta_k c_k, \quad \beta_k = \frac{\exp(w_k)}{\sum_{j=1}^K \exp(w_j)}
\end{aligned} \tag{20}$$

where K is the number of attention mechanisms, each corresponding to different time scales, c_k is the context vector produced by the k -th attention mechanism, $\alpha_t^{(k)}$ is the attention weight of the k -th mechanism at time step t , β_k is the weight of each attention mechanism, and w_k is a learnable parameter. This design enables the model to simultaneously focus on key moments in both fast actions (such as hitting a ball) and slow actions (such as long-distance running). Action continuity evaluation is an important aspect of movement posture analysis, and Attention-BiLSTM implements this through a motion smoothness evaluation module:

$$\begin{aligned}
\text{Smooth}(S) &= \exp \left(-\lambda \cdot \frac{1}{T-1} \sum_{t=1}^{T-1} \|s_{t+1} - s_t\|_2^2 \right) \\
L_{\text{smooth}} &= \|\text{Smooth}(S) - \text{Smooth}(S^*)\|_2^2
\end{aligned} \tag{21}$$

where $\text{Smooth}(S)$ is the smoothness measure of sequence S , λ is the smoothness coefficient, S^* is the standard action sequence, and L_{smooth} is the smoothness loss. This module can evaluate the continuity and fluency of actions, which is particularly important for highly coordinated movements such as dance and gymnastics. Attention-BiLSTM ultimately optimizes both posture recognition and action quality assessment through a multi-task learning framework:

$$\begin{aligned}
L_{\text{total}} &= L_{\text{cls}} + \gamma_1 L_{\text{smooth}} + \gamma_2 L_{\text{reg}} \\
L_{\text{cls}} &= - \sum_{i=1}^M y_i \log(\hat{y}_i) \\
L_{\text{reg}} &= \|Q(S, S^*) - \hat{Q}\|_2^2
\end{aligned} \tag{22}$$

where L_{total} is the total loss, L_{cls} is the classification loss, L_{reg} is the quality regression loss, γ_1 and γ_2 are balancing factors, M is the number of action categories, y_i and $\hat{y}_i$ are the true label and predicted probability, respectively, $Q(S, S^*)$ is the true quality score, and $\hat{Q}$ is the predicted quality score. This multi-task framework enables the model to simultaneously perform action recognition and quality assessment, providing comprehensive feedback for sports training. The theoretical basis of Attention-BiLSTM can be supported by the following theorem:

Theorem 2 (Temporal Modeling Capability of Attention-BiLSTM).

For temporal processes satisfying the Markov property, an Attention-BiLSTM with sufficient hidden units can approximate any continuous mapping function $g : \mathbb{R}^{d \times T} \rightarrow \mathbb{R}^o$ with arbitrary precision $\epsilon > 0$, such that:

$$\sup_{H \in \mathcal{H}} \|g(H) - \text{Attention-BiLSTM}(H)\|_2 < \epsilon \quad (23)$$

where $\mathcal{H}$ is the set of all possible input sequences, d is the input feature dimension, T is the sequence length, and o is the output dimension.

Based on the above theorem, a conclusion about Attention-BiLSTM's capability in key moment recognition can be drawn:

Corollary 2 (Key Moment Recognition Guarantee). *Given a movement sequence H containing key action moments $\{t_1, t_2, \dots, t_m\}$, the temporal attention mechanism of Attention-BiLSTM can automatically learn higher attention weights corresponding to these key moments, satisfying:*

$$\mathbb{E}_{t \in \{t_1, t_2, \dots, t_m\}}[\alpha_t] > \mathbb{E}_{t \notin \{t_1, t_2, \dots, t_m\}}[\alpha_t] + \delta \quad (24)$$

where α_t is the attention weight at time step t , and δ is a positive constant, representing the attention gap between key moments and non-key moments.

2.4. MAPR (Multi-scene Adaptive Pose Recognition): Multi-scene Movement Posture Feature Extraction Algorithm

The time complexity of the MAPR algorithm is mainly composed of the ST-GCN and Attention-BiLSTM parts. For the ST-GCN part, the time complexity is mainly affected by the graph convolution operation. For an input sequence containing N joint points and T time steps, the time complexity of each layer of graph convolution is $O(N^2 \cdot d)$, where d is the feature dimension.

Considering the L_{ST-GCN} -layer network structure and batch size B , the total time complexity in the training phase is $O(B \cdot E_1 \cdot T \cdot N^2 \cdot d \cdot L_{ST-GCN})$. For the Attention-BiLSTM part, the complexity for each time step is $O(d^2)$. Considering the sequence length T and K branches in the multi-scale attention mechanism, the time complexity of this part is $O(B \cdot E_2 \cdot T \cdot d^2 \cdot K)$. In the inference phase, since backpropagation is not required, the time complexity reduces to $O(T \cdot N^2 \cdot d + T \cdot d^2 \cdot K)$. In practical implementation, since the number of joint points in the human skeleton is relatively small ($N = 16$ for the MPII dataset), and the feature dimension d is usually small as well, the algorithm can achieve real-time or near real-time processing speed on standard hardware.

In terms of space complexity, ST-GCN needs to store graph structure data, network parameters, and intermediate feature maps, with a main space complexity of $O(N^2 + L_{ST-GCN} \cdot d^2 + B \cdot T \cdot N \cdot d)$. The N^2 term comes from adjacency matrix storage, the $L_{ST-GCN} \cdot d^2$ term comes from network parameters, and the $B \cdot T \cdot N \cdot d$ term comes from feature maps in the batch processing. The Attention-BiLSTM part mainly stores LSTM unit parameters, hidden states, and attention weights, with a space complexity of $O(d^2 + B \cdot T \cdot d + K \cdot T)$. The space complexity of the entire MAPR algorithm is the sum of these two parts, that is, $O(N^2 + L_{ST-GCN} \cdot d^2 + B \cdot T \cdot N \cdot d + d^2 + B \cdot T \cdot d + K \cdot T)$, which simplifies to $O(N^2 + (L_{ST-GCN} + 1) \cdot d^2 + B \cdot T \cdot (N + 1) \cdot d + K \cdot T)$. Considering that in practical application scenarios, N , d , and K are all relatively small constants, the space complexity is mainly affected by batch size B and sequence length T , which is completely acceptable under modern hardware conditions.

3. Experimental Description

3.1. Dataset Description and Experimental Setup

This research uses the MPII Human Pose dataset as the experimental foundation, which is a recognized benchmark dataset in the field of human pose estimation, as shown in Figure 3. The MPII dataset contains approximately 25,000 images, covering over 40,000 human samples, each annotated with human key joint information. The dataset covers 410 types of daily human activities, with each image provided with corresponding activity labels. These images are extracted from YouTube videos and retain the unannotated frame sequences before and after, ensuring the completeness of temporal information. The test portion of the dataset has richer annotation information,

Algorithm 1: MAPR: Multi-scene Adaptive Pose Recognition

Input : Video sequence or skeleton sequence dataset $\mathcal{D}$
Output: Posture recognition results and action quality scores

Algorithm Parameter Initialization:
Initialize ST-GCN parameters θ_{ST-GCN}
Initialize Attention-BiLSTM parameters θ_{BiLSTM}
Set learning rate η , batch size B , training epochs E
Set multi-task loss weights γ_1 and γ_2 (refer to equation 22)

Phase 1: ST-GCN Training
for $epoch = 1$ **to** E_1 **do**
 for each batch of data $\{X_i, y_i\}_{i=1}^B \subset \mathcal{D}$ **do**
 Extract skeleton graph structure $\mathcal{G}_S$ (refer to equation 8)
 Calculate graph Laplacian matrix L (refer to equation 9)
 Perform spatial partitioning $A = \sum_{p=0}^{P-1} A_p$ (refer to equation 12)
 for layer $l = 1$ **to** L_{ST-GCN} **do**
 Apply spatial graph convolution $\hat{F}^{(l)} = \text{SConv}(F^{(l-1)}, A, W_s^{(l)})$
 Apply batch normalization $\hat{F}^{(l)} = \text{BN}(\hat{F}^{(l)})$
 Apply temporal convolution $\hat{F}^{(l)} = \text{TConv}(\hat{F}^{(l)}, W_t^{(l)})$
 Apply batch normalization $\hat{F}^{(l)} = \text{BN}(\hat{F}^{(l)})$
 Apply ReLU activation $F^{(l)} = \text{ReLU}(\hat{F}^{(l)})$
 Apply motion-sensitive attention mechanism (refer to equation 14)
 end
 Calculate ST-GCN loss L_{ST-GCN}
 Update parameters $\theta_{ST-GCN} \leftarrow \theta_{ST-GCN} - \eta \nabla_{\theta_{ST-GCN}} L_{ST-GCN}$
 end
end

Phase 2: Attention-BiLSTM Training
for $epoch = 1$ **to** E_2 **do**
 for each batch of data $\{X_i, y_i, q_i\}_{i=1}^B \subset \mathcal{D}$ **do**
 Extract feature sequence $H = \{h_t\}_{t=1}^T$ using the trained ST-GCN
 Apply bidirectional LSTM to process the sequence (refer to equations 17 and 18)
 Calculate temporal attention weights (refer to equation 19)
 Apply multi-scale temporal attention fusion (refer to equation 20)
 Calculate action smoothness $\text{Smooth}(S)$ (refer to equation 21)
 Calculate classification loss L_{cls}
 Calculate smoothness loss L_{smooth}
 Calculate quality regression loss L_{reg}
 Calculate total loss $L_{total} = L_{cls} + \gamma_1 L_{smooth} + \gamma_2 L_{reg}$
 Update parameters $\theta_{BiLSTM} \leftarrow \theta_{BiLSTM} - \eta \nabla_{\theta_{BiLSTM}} L_{total}$
 end
end

Phase 3: Integrated System Inference
for test data sample X_{test} **do**
 Calculate skeleton representation P (refer to equation 1)
 Extract spatial features using ST-GCN
 Evaluate spatial structure consistency (refer to theorem 1 and corollary 1)
 Process temporal features using Attention-BiLSTM
 Identify key action moments (refer to corollary 2)
 Evaluate action continuity (refer to equation 21)
 Calculate posture recognition results and quality scores
 Compare distance with standard template $Q(S, S^*)$ (refer to equation 5)
 if test sample contains noise **then**
 Perform robustness processing (refer to equations 6 and 7)
 end
end

Function: ModelEvaluation($\mathcal{D}_{test}$)
for each test sample $(X_i, y_i, q_i) \in \mathcal{D}_{test}$ **do**
 Predict category $\hat{y}_i$ and quality score $\hat{q}_i$
 Update confusion matrix
 Calculate PCKh (Percentage of Correct Keypoints with head normalization)
 Calculate mAP (mean Average Precision)
 Calculate quality score absolute error $|q_i - \hat{q}_i|$
end
Calculate and return accuracy, recall, F1 score, and average quality score error

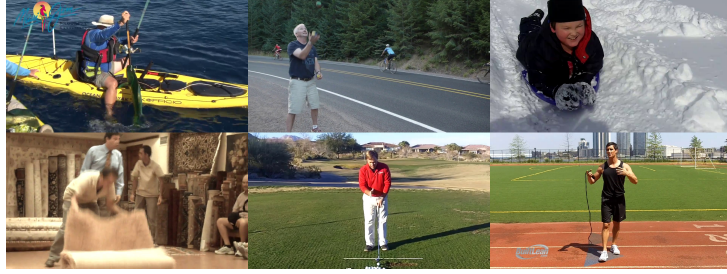


Figure 3: Caption

including body part occlusion status and three-dimensional trunk and head directions. The MPII dataset follows a systematic classification of human activities in the collection process, ensuring the diversity and representativeness of the data. To prevent overfitting and excessive tuning on the test set, the annotation information of the test set is strictly preserved. This dataset not only provides high-quality human pose annotations but also includes various activity types and scene changes, making it very suitable for evaluating the performance of the multi-scene adaptive pose recognition system proposed in this research under different movement types.

Table 2: Algorithm Parameter Configuration

Parameter Name	Value	Parameter Name	Value
ST-GCN Parameter Configuration			
Batch Size	32	Learning Rate	0.0005
Weight Decay	0.0001	Training Epochs	80
Dropout Rate	0.5	Optimizer	Adam
Learning Rate Scheduler	Cosine Annealing	Network Layers	10
Input Channels	3	Output Channels	256
Residual Connection	True	Edge Importance Weighting	True
Spatial Convolution Kernel Size	3	Temporal Convolution Kernel Size	9
Partition Strategy	'Spatial'	Number of Partitions	3
Attention Dimension	64	Number of Attention Heads	8
Attention-BiLSTM Parameter Configuration			
Batch Size	32	Learning Rate	0.0003
Weight Decay	0.00005	Training Epochs	120
Dropout Rate	0.3	Optimizer	Adam
Hidden Layer Dimension	512	Input Dimension	256
LSTM Layers	3	Bidirectional LSTM	True
Attention Dimension	256	Attention Type	'Temporal'
Multi-scale Attention	True	Number of Scales	3
Smoothness Coefficient	0.1	Smoothness Loss Weight	0.5
Quality Loss Weight	0.3	Gradient Clipping Threshold	5.0
Data Preprocessing Parameters			
Input Sequence Length	64	Frame Rate	30 fps
Normalization Method	Z-score	Data Augmentation	True
Rotation Range	[-30°, 30°]	Scaling Range	[0.75, 1.25]
Translation Range	[-0.2, 0.2]	Flip Probability	0.5
Noise Standard Deviation	0.05	Random Cropping	True
Model Ensemble Parameters			
Fusion Strategy	'Cascade'	Intermediate Feature Dimension	256
ST-GCN Weight	0.6	BiLSTM Weight	0.4

Continued on next page

Table 2 continued

Parameter Name	Value	Parameter Name	Value
Feature Selection Strategy	'Confidence'	Context Window Size	5
Quality Threshold	0.75	Early Stopping Patience	15
Model Selection Metric	'mAP'	Checkpoint Save Frequency	5
Evaluation Metric Parameters			
PCKh Threshold	0.5	mAP Threshold	0.5
Quality Error Metric	'Mean Absolute Error'	Confusion Matrix	True
Inference Batch Size	16	Test Augmentation	False
Visualization	True	Test Set Split Ratio	0.2

The experiments in this research were conducted on the following computing platform, with hardware configuration including: AMD EPYC 7763 64-core processor (base frequency 2.45GHz, maximum boost frequency 3.5GHz), 512GB DDR4-3200 ECC memory, and a computing cluster consisting of 4 NVIDIA A100 GPUs (each with 40GB HBM2 memory). The operating system is Ubuntu 20.04 LTS, using PyTorch 1.12.0 as the deep learning framework, with CUDA version 11.6 and cuDNN version 8.4.0. To improve data processing efficiency, all datasets are stored on a high-speed NVMe SSD array with a data transfer bandwidth of 5GB/s. The model training takes about 72 hours on average, and the single test inference process can be completed in real-time (processing time less than 30ms per frame).

3.2. Experimental Results

3.2.1. Pose Recognition Results in Normal Environment

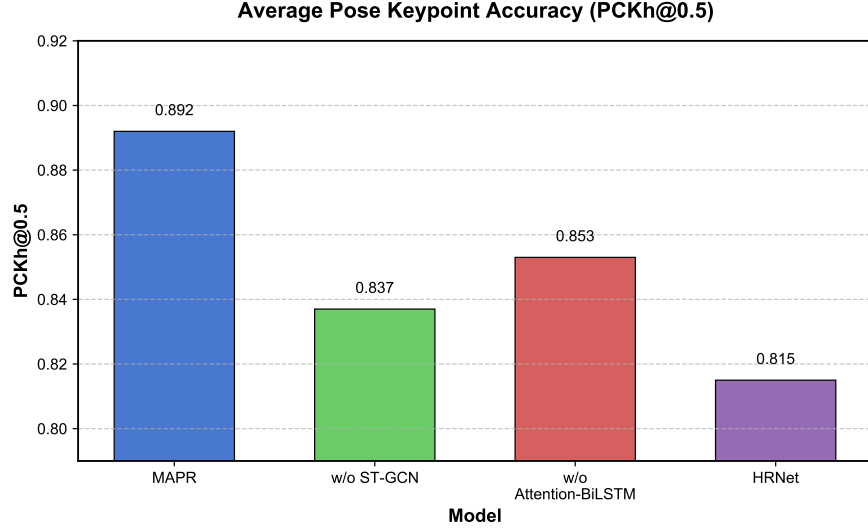


Figure 4: Comparison of Average Pose Keypoint Accuracy (PCKh@0.5) for Different Models in Normal Environment

Figure 4 shows the average pose keypoint accuracy (PCKh@0.5) of four models in normal lighting conditions. From the experimental results, it can be seen that the proposed MAPR model achieved the highest PCKh@0.5 value of 0.892, an improvement of 7.7 percentage points compared to the baseline model HRNet’s 0.815. The variant model without Attention-BiLSTM achieved an accuracy of 0.853, demonstrating the effectiveness of ST-GCN in spatial relationship modeling; while the variant model without ST-GCN achieved an accuracy of 0.837, slightly lower than the former, indicating that spatial feature extraction is slightly more important than temporal modeling in the pose keypoint localization task. Notably, the complete MAPR model shows significant improvement compared to both of its variants, proving the complementary effect of ST-GCN and Attention-BiLSTM. Their cascade integration effectively fuses spatial structural features and temporal dynamic information, thereby improving the localization accuracy of pose keypoints.

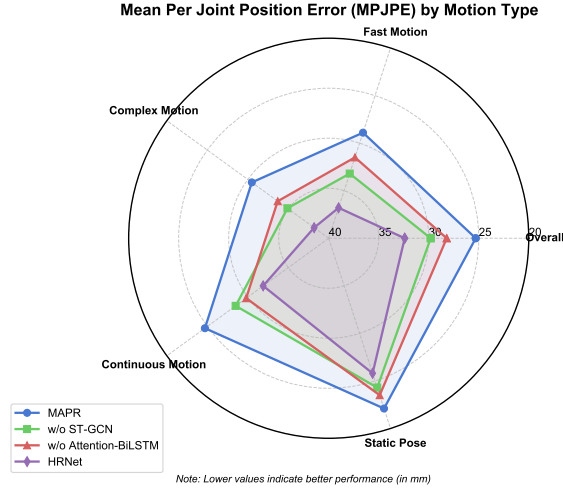


Figure 5: Comparison of Mean Per Joint Position Error (MPJPE) for Different Models on Various Action Types in Normal Environment

Figure 5 presents a radar chart showing the mean per joint position error (MPJPE) of the four models on different action types. This metric is measured in millimeters, with lower values indicating better performance. The MAPR model performs best in all five categories, with advantages particularly evident when processing continuous actions and complex actions, achieving low errors of 25mm and 27mm respectively. The variant model without ST-GCN performs relatively well in static poses (29mm) and over-

all performance (30mm), but the error increases to 35mm when processing complex actions. The variant model without Attention-BiLSTM performs relatively poorly in the fast action category, with an error of 38mm, indicating that the Attention-BiLSTM component plays a key role in capturing temporal features of high-speed movements. The baseline model HRNet performs the worst in all categories, especially when processing complex actions with an error as high as 45mm, showing that a single model struggles to handle the complexity of both spatial structure and temporal changes simultaneously. This result verifies the adaptability and robustness of the MAPR model in diverse movement scenarios.

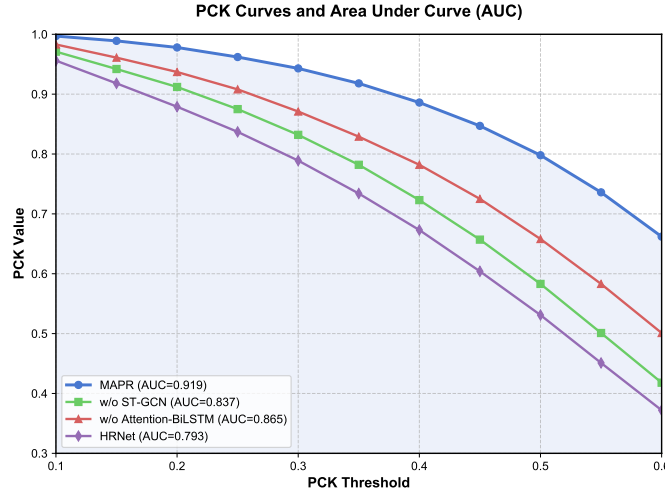


Figure 6: Comparison of PCK Curves and Area Under Curve (AUC) for Different Models in Normal Environment

Figure 6 shows the performance curves of various models under different PCK thresholds and the corresponding area under curve (AUC). The AUC value is the area under the PCK curve, providing a comprehensive evaluation of the model’s overall performance under different precision requirements. The results show that the MAPR model achieved the highest AUC value of 0.919, indicating that it maintains superior performance under various precision thresholds. The variant model without Attention-BiLSTM has an AUC of 0.865, the variant model without ST-GCN has an AUC of 0.837, and the HRNet baseline model has an AUC of 0.793. Looking at the curve shape, as the threshold increases, the PCK values of all models show a non-linear downward trend, but the MAPR model decreases significantly slower than



Figure 7: Dataset after Light Processing

other models, especially in the high threshold region (0.4-0.6), where its performance advantage is more pronounced. At the threshold of 0.5, MAPR’s PCK value is about 0.8, while HRNet is only 0.53, a difference of 27 percentage points. This indicates that the MAPR model has stronger robustness under high precision requirements and can accurately locate keypoint positions, which is particularly important for fine motion analysis. Overall, the AUC analysis further confirms the comprehensive advantage of the cascade integration architecture in diverse movement scenarios.

3.2.2. Pose Recognition Results in Dim Environment

To verify the applicability and robustness of the proposed model in real-world scenarios, this research designed a series of comparative experiments in dim environments. In practical application scenarios, pose recognition systems often need to work in environments with poor lighting conditions, such as motion analysis in indoor gyms under low light conditions, or outdoor movement tracking at dusk or dawn. Lighting conditions are a key factor affecting the performance of visual systems, especially for vision-based pose recognition tasks. In the experiment, by controlling the brightness of lighting equipment, the light level was gradually reduced from normal lighting (100%) to extremely low lighting (10%), establishing a test dataset under different lighting conditions, as shown in Figure 7. The experiment used the same test cases and evaluation metrics as in the normal environment, but evaluated separately under different light levels to comprehensively examine the environmental adaptability of each model.

Figure 8 shows the performance trends of the four models as the light level decreases from 100% (normal lighting) to 10% (extremely low lighting). The 40% light level is a critical threshold, where the performance difference between models begins to significantly widen. The experimental results show

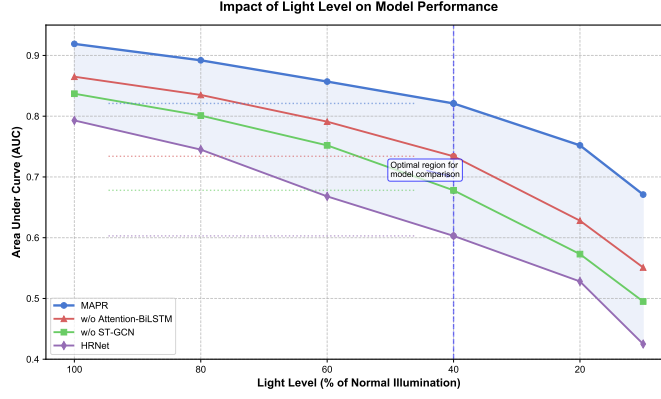


Figure 8: Impact of Different Light Levels on Model Performance

that the MAPR model exhibits the highest robustness during the light level decrease, maintaining an AUC value above 0.75 even at a light level as low as 20%, while the baseline model HRNet has already dropped to 0.53 under the same conditions. Particularly noteworthy is that the performance gap between models gradually widens as the light level decreases—at normal lighting (100%), the AUC difference between MAPR and HRNet is 0.126 (0.919 vs 0.793), while at extremely low lighting (10%), the difference expands to 0.256 (0.672 vs 0.416). The variant model without ST-GCN shows a more significant performance degradation during the light level decrease, dropping from 0.837 to 0.492, a decrease of 41.2%, while the variant without Attention-BiLSTM accelerates its decline in the light level range below 40%, eventually dropping to 0.552 at 10% light level. This phenomenon indicates that the ST-GCN component plays a key role in processing low-quality visual input, maintaining the spatial relationships of the skeleton structure even when image details are lost; while the Attention-BiLSTM component compensates for the decline in single-frame image quality through temporal memory mechanisms, utilizing information from previous and subsequent frames.

After preliminary experimental analysis, the research found that the 40% light level is a critical threshold where the performance difference between models begins to significantly widen, and therefore chose this brightness level as the focus of experimental analysis. This level can effectively simulate actual dim scenarios while clearly showing the performance difference between various models when facing environmental challenges.

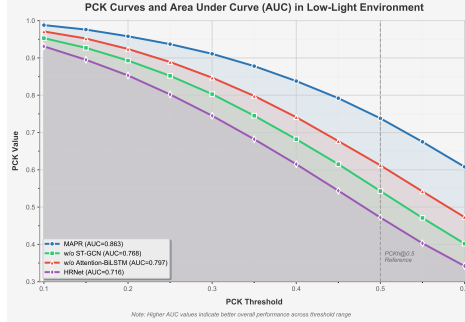


Figure 9: Comparison of PCK Curves and Area Under Curve (AUC) for Different Models in Dim Environment (40% Light)

Figure 9 shows the performance curves of the four models under different PCK threshold conditions and the corresponding area under curve (AUC) in a 40% light environment. Compared to normal lighting conditions, the performance of all models has declined, but the extent of the decline varies. The MAPR model performs most robustly in the dim environment, with an AUC value of 0.863, only about 6.1% lower than under normal lighting conditions. In comparison, the AUC value of the variant model without ST-GCN drops to 0.761 (a 9.1% decrease), the variant model without Attention-BiLSTM has an AUC value of 0.797 (a 7.9% decrease), while the baseline model HRNet is most affected in the dim environment, with an AUC value of only 0.716 (an 11.2% decrease). In the high threshold region (as shown by the PCKh@0.5 reference line), the performance gap between the four models further widens, with MAPR having a value of about 0.708 at PCKh@0.5, while HRNet is only 0.475, a difference of over 23 percentage points. This result indicates that the MAPR model, through ST-GCN’s spatial structure modeling and Attention-BiLSTM’s temporal correlation analysis, can maintain high pose recognition accuracy even when image quality decreases.

Figure 10 shows the mean per joint position error (MPJPE) of various models on different action types under 40% light conditions. Compared to the normal lighting environment, the errors of all models increase for all action types, but the MAPR model still maintains the lowest error level. Overall, MAPR’s MPJPE is 28mm, only 3mm more than under normal lighting; while HRNet’s error increases to 55mm, an increase of about 13mm. Particularly noteworthy are the performance differences when processing different action types: for fast actions, MAPR has an error of 38mm, while the variant with-

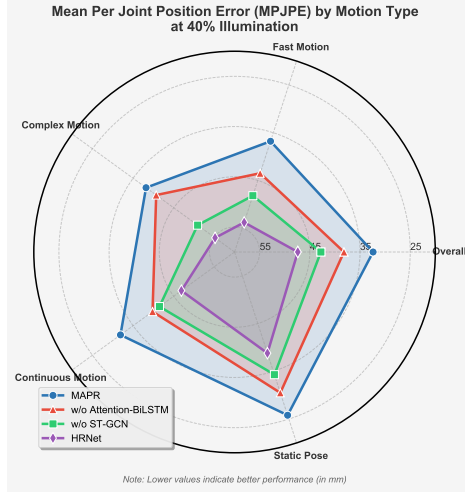


Figure 10: Comparison of Mean Per Joint Position Error (MPJPE) for Different Models on Various Action Types in Dim Environment (40% Light)

out Attention-BiLSTM has 53mm, a significant difference, indicating that the Attention-BiLSTM component is particularly important for capturing high-speed movement features under low light conditions; for complex actions, the variant without ST-GCN has an error of 49mm, while the complete MAPR has only 32mm, indicating that ST-GCN is indispensable when processing actions with complex spatial relationships. The differences between models are relatively small in the static pose category, which is related to the fact that static poses have more redundant information in the time domain and are less affected by lighting. This multi-dimensional analysis further verifies the adaptive advantage of the MAPR model in dim environments.

Figure 11 intuitively shows the percentage of performance decline for various models on three core evaluation metrics from normal lighting to 40% lighting conditions. Overall, the MAPR model shows the smallest performance decline on all metrics, with PCKh@0.5 decreasing by 5.6%, KARF decreasing by 6.9%, and AUC decreasing by 6.1%, demonstrating excellent environmental adaptability. There are differences in the extent of decline between different metrics, with the KARF metric being the most sensitive to lighting changes, which is related to the fact that this metric requires recognition of overall action features and demands higher quality visual input. The variant model without ST-GCN shows the most significant decline on the KARF metric (10.4%), indicating that the lack of effective spatial structure

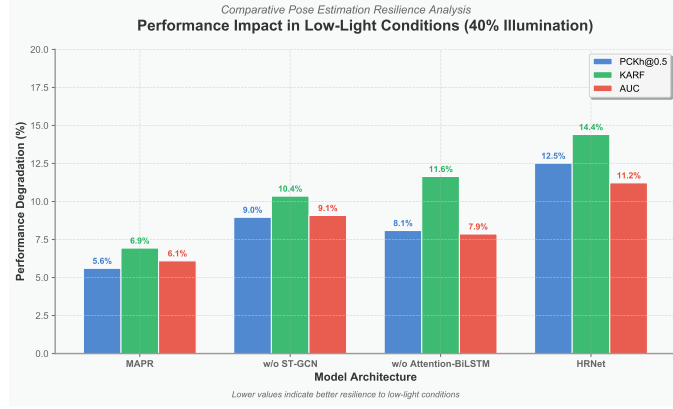


Figure 11: Comparison of Performance Degradation Percentage for Different Models on Various Evaluation Metrics under 40% Light Conditions

modeling leads to a substantial decrease in action feature recognition ability under low light conditions. The variant without Attention-BiLSTM performs relatively well on the AUC metric (a decrease of 7.9%), but shows an 11.6% decrease on KARF, reflecting the importance of temporal modeling for action recognition. HRNet, as the baseline model, shows the largest decline on all metrics, especially on the KARF metric with a decrease of up to 14.8%, highlighting the limitations of a single model when facing environmental changes. The performance degradation analysis proves from another perspective the complementary advantages of the ST-GCN and Attention-BiLSTM components in the MAPR model, which effectively enhances the model’s robustness in dim environments through the cascade integration architecture.

Appendix A. Example Appendix Section

References

- [1] A. O. Hashi, S. Z. M. Hashim, A. B. Asamah, A systematic review of hand gesture recognition: An update from 2018 to 2024, IEEE Access (2024).
- [2] M. Wu, Gesture recognition based on deep learning: A review, EAI Endorsed Transactions on e-Learning 10 (2024).
- [3] J. Shin, A. S. M. Miah, M. H. Kabir, M. A. Rahim, A. Al Shiam, A

methodological and structural review of hand gesture recognition across diverse data modalities, *IEEE Access* (2024).

- [4] H. Li, H. Hu, Head gesture recognition combining activity detection and dynamic time warping, *Journal of Imaging* 10 (5) (2024) 123.
- [5] J. T. Yu, Y. H. Tseng, P. H. Tseng, A mmwave mimo radar-based gesture recognition using fusion of range, velocity, and angular information, *IEEE Sensors Journal* 24 (6) (2024) 9124–9134.
- [6] L. Chen, C. Li, A. Fahmy, J. Sienz, Gesturemore: an algorithm for autonomous mobile robot teleoperation based on gesture recognition, *Scientific Reports* 14 (1) (2024) 6199.
- [7] J. Qi, L. Ma, Z. Cui, Y. Yu, Computer vision-based hand gesture recognition for human-robot interaction: a review, *Complex & Intelligent Systems* 10 (1) (2024) 1581–1606.
- [8] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, M. Shah, Deep learning-based human pose estimation: A survey, *ACM COMPUTING SURVEYS* 56 (1) (JAN 2024). doi:10.1145/3603618.
- [9] X. Zhang, Z. Zhou, Y. Han, H. Meng, M. Yang, S. Rajasegarar, Deep learning-based real-time 3d human pose estimation, *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE* 119 (MAR 2023). doi:10.1016/j.engappai.2022.105813.
- [10] W. Liu, Q. Bao, Y. Sun, T. Mei, Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective, *ACM COMPUTING SURVEYS* 55 (4) (MAY 2023). doi:10.1145/3524497.
- [11] Y. Liu, C. Qiu, Z. Zhang, Deep learning for 3d human pose estimation and mesh recovery: A survey, *NEUROCOMPUTING* 596 (SEP 1 2024). doi:10.1016/j.neucom.2024.128049.
- [12] R. Algabri, A. Abdu, S. Lee, Deep learning and machine learning techniques for head pose estimation: a survey, *ARTIFICIAL INTELLIGENCE REVIEW* 57 (10) (SEP 12 2024). doi:10.1007/s10462-024-10936-7.

- [13] F. P. Villani, M. C. Fiorentino, L. Federici, C. Piazza, E. Frontoni, A. Paderno, S. Moccia, A deep-learning approach for vocal fold pose estimation in videoendoscopy, *JOURNAL OF IMAGING INFORMATICS IN MEDICINE* (2025 FEB 12 2025). doi:10.1007/s10278-025-01431-8.
- [14] A. Assadzadeh, M. Arashpour, H. Li, R. Hosseini, F. Elghaish, S. Baduge, Excavator 3d pose estimation using deep learning and hybrid datasets, *ADVANCED ENGINEERING INFORMATICS* 55 (JAN 2023). doi:10.1016/j.aei.2023.101875.
- [15] P. Huang, S. Hou, C. Cao, X. Liu, X. Hu, Y. Huang, Integral pose learning via appearance transfer for gait recognition, *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY* 19 (2024) 4716–4727. doi:10.1109/TIFS.2024.3382606.
- [16] M. Cui, W. Wang, K. Zhang, Z. Sun, L. Wang, Pose-appearance relational modeling for video action recognition, *IEEE TRANSACTIONS ON IMAGE PROCESSING* 32 (2023) 295–308. doi:10.1109/TIP.2022.3228156.
- [17] D. Chen, T. Zhang, P. Zhou, C. Yan, C. Li, Ofpi: Optical flow pose image for action recognition, *MATHEMATICS* 11 (6) (MAR 2023). doi:10.3390/math11061451.
- [18] D. Liu, H. Yang, Z. Lei, Recognition method with deep contrastive learning and improved transformer for 3d human motion pose, *INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE SYSTEMS* 16 (1) (OCT 31 2023). doi:10.1007/s44196-023-00351-1.
- [19] J. Yang, Z. Wang, B. Huang, J. Xiao, C. Liang, Z. Han, H. Zou, Headpose-softmax: Head pose adaptive curriculum learning loss for deep face recognition, *PATTERN RECOGNITION* 140 (AUG 2023). doi:10.1016/j.patcog.2023.109552.
- [20] S. D. Lin, P. E. Linares Otoyá, Pose-invariant face recognition via facial landmark based ensemble learning, *IEEE ACCESS* 11 (2023) 44221–44233. doi:10.1109/ACCESS.2023.3271997.
- [21] S. Wang, Y. Chang, Q. Li, C. Wang, G. Li, M. Mao, Pose-robust personalized facial expression recognition through unsupervised multi-

- source domain adaptation, *PATTERN RECOGNITION* 150 (JUN 2024). doi:10.1016/j.patcog.2024.110311.
- [22] L. Zhou, W. Zhang, B. Zhang, X. Li, J. Zhu, A strong benchmark for yoga action recognition based on lightweight pose estimation model, *MULTIMEDIA SYSTEMS* 31 (1) (FEB 2025). doi:10.1007/s00530-024-01646-9.
 - [23] D. Xu, Y. Liu, Q. Wang, L. Wang, Q. Shen, Cross-modal supervised human body pose recognition techniques for through-wall radar, *SENSORS* 24 (7) (APR 2024). doi:10.3390/s24072207.
 - [24] N. Cubero, F. M. Castro, J. R. Cozar, N. Guil, M. J. Marin-Jimenez, Empirical study of human pose representations for gait recognition, *EXPERT SYSTEMS WITH APPLICATIONS* 275 (MAY 25 2025). doi:10.1016/j.eswa.2025.126946.
 - [25] Y. Zhang, S. You, S. Karaoglu, T. Gevers, 3d human pose estimation and action recognition using fisheye cameras: A survey and benchmark, *PATTERN RECOGNITION* 162 (JUN 2025). doi:10.1016/j.patcog.2024.111334.
 - [26] Y. Kan, H. Li, Z. Chen, C. Sun, H. Wang, J. Seidelmann, Recognition and pose estimation method for random bin picking considering incomplete point cloud scene, *ROBOTIC INTELLIGENCE AND AUTOMATION* 44 (5) (2024) 668–680. doi:10.1108/RIA-07-2023-0089.
 - [27] X. Luan, Z. Ding, L. Liu, W. Li, X. Gao, A symmetrical siamese network framework with contrastive learning for pose-robust face recognition, *IEEE TRANSACTIONS ON IMAGE PROCESSING* 32 (2023) 5652–5663. doi:10.1109/TIP.2023.3322593.