

Wushu 3D Pose Estimation and Transformer-GRU Fusion Recognition Method

Yufeng Ma

Associate Professor, Special Police College, Nanjing Police University, Nanjing, 210023, China, Email:
Mayufeng20251114@163.com

Project Management

Received February 26, 2026; revised April 7, 2026; accepted August 30, 2026
Available online September 20, 2026

Abstract: In order to address the limitations of traditional Wushu 3D posture recognition, such as low efficiency and accuracy, the study extracts key joint points of the human body using the Mediapipe technique. The Transformer is utilized to capture dynamic correlations and long-term dependencies in martial arts action sequences. The Bidirectional Gated Recurrent Unit (BiGRU) establishes global spatiotemporal dependencies, while the Graph Convolutional Network (GCN) constructs the spatial structure of martial arts postures. On this basis, a three-dimensional pose recognition model for Wushu is constructed by integrating Depth-Separable Convolution (DSC). The results show that the proposed algorithm achieves optimal adaptation in the shortest time for Wushu 3D posture recognition, with recognition accuracy reaching 97.01% and a loss value of 0.04. The proposed model has an average error of only 1.99 mm per joint in Wushu 3D posture recognition, and the key point detection accuracy reaches 94.32%, both of which are better than those of the comparison models. Meanwhile, the recognition accuracies for the two martial arts scenarios, Taijiquan and Wing Chun, are 95.96% and 95.87%, respectively. Taken together, the model shows superior accuracy in recognizing 3D martial arts postures and demonstrates good operational stability, providing technical support for related fields.

Keywords: Wushu; 3D pose estimation; transformer; Bidirectional Gated Recurrent Unit (BiGRU); Graph Convolutional Network (GCN); mediapipe.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-262

1. Introduction

Martial arts is a combat system that combines highly technical offensive and defensive strategies with profound cultural and philosophical connotations, and its movements involve complex physical and mental coordination. Accurate recognition and quantitative analysis of martial arts techniques can provide real-time feedback and objective scoring for athletes' scientific training (Chen, 2025). However, traditional martial arts posture recognition relies heavily on apprenticeship experience and subjective observation, which leads to time-consuming, laborious manual analysis and difficulty achieving real-time feedback (Li and Wu, 2023; Roccliffe et al., 2024). Therefore, there is an urgent need for efficient, accurate technology for recognizing martial arts postures.

With the rapid development of computer technology, intelligent algorithms have become a research hotspot in many fields in recent years. Among them, Transformer uses transformers to process feature data, enhancing parallel computing capabilities (Zhang et al., 2024). Pradani and Suadaa (2023) proposed an automatic scoring model, but it limited its potential to improve on a Transformer for evaluating paper exam answers. The model extracts features through weighted techniques of information retrieval and text mining, calculates semantic similarity using cosine similarity, and finally accurately evaluates paper scores. The Gated Recurrent Unit (GRU) effectively captures long- and short-term dependencies in temporal data by controlling the flow and forgetting of sequence information between time steps (Ashtari-Majlan and Masip, 2025; Chauhan and Kumar, 2024). In order to improve the channel estimation performance of bit-to-analog converters, Helmy et al. (2023) proposed a channel matrix model that combines long short-term memory networks and GRU. Splitting the received signal into multiple subsystems for sequential training effectively leverages high-signal-to-noise-ratio data and eliminates the gradient-vanishing problem in traditional training. Transformers can achieve efficient parallel computation through self-attention mechanisms, but they have high computational complexity when processing extremely long sequences. The GRU structure is relatively lightweight and has fewer parameters than Transformers, but it cannot process the entire sequence in parallel, limiting its potential to improve training speed.

Vision-based 3D human pose estimation technology offers new approaches to intelligent Wushu analysis, and many scholars have conducted research in this area. Kishore and Kumar (2025) constructed a lightweight real-data embedding network that recognizes video data using four fully connected layers, extracts 3D skeletal joint data from videos, and establishes the relationship between 3D pose estimation and motion capture, thereby producing pose vectors. Xu et al. (2023) proposed an abnormal human pose recognition method that fuses multi-view cross information and confidence measurement. The approach extracts skeletal information using OpenPose and employs YOLOv5 to detect abnormal poses, such as falls, to improve the efficiency and accuracy of abnormal pose detection. Myung et al. (2024) discussed the neglect of intra-class variety and redundancy in skeletal action recognition, proposing a novel model with adaptive sampling-point learning on spatiotemporal graphs to emphasize key points and a multi-branch network to better fuse joint and skeleton structures, thereby improving model discriminability over key points. Existing studies fall into two categories: improving real-time tracking efficiency and enhancing model robustness, but they are not well balanced with respect to adaptability, efficiency, and accuracy for related actions in Wushu.

Therefore, to address the shortcomings of the current study, the study utilizes Mediapipe to extract static joint features, Transformer to capture dynamic associations and long-term dependencies in Wushu action sequences, and BiGRU to establish global spatio-temporal dependencies. At the same time, GCN is introduced to construct the spatial structure of martial arts gestures. On this basis, DSC is integrated to construct a three-dimensional posture recognition model for martial arts. It is expected that the proposed model can accurately recognize 3D postures in martial arts and provide data support for related fields. The research innovations and contributions are listed below:

1. Mediapipe is used to extract joint feature nodes and combine the Transformer and BiGRU algorithms to achieve the extraction and enhancement of martial arts 3D poses. This combination effectively addresses the limitation that a single model cannot simultaneously account for both global spatiotemporal dependence and local continuous dynamics in martial arts movements, and enables accurate extraction and dynamic enhancement of 3D martial arts posture.
2. Introduce GCN to highlight the core key posture joints in martial arts movements and establish the spatial relationship of the human skeleton. This module highlights the relationships among key postural joints in martial arts movements, enhances the model's understanding of movement structure, and thus improves the accuracy of recognizing complex martial arts moves.
3. Combining DSC for lightweight design, significantly reducing the number of model parameters and computational costs while maintaining high recognition accuracy. The final research integrates the strengths of multiple modules to develop a 3D pose recognition model for martial arts, providing practical technical references for applications such as intelligent training and real-time feedback.

2. Methodology

2.1. Wushu 3D Pose Estimation with Transformer and BiGRU

Martial arts 3D pose estimation can capture human motion trajectories and has become an important research direction in sports (Mokayed et al., 2023). Traditional methods rely heavily on optical markers or inertial sensors, which suffer from high latency and insufficient accuracy (Guo et al., 2024). Mediapipe can collect real-time motion videos on the device and extract human joint nodes, converting image understanding into structured joint data (Sánchez-Brizuela et al., 2023; Vu et al., 2024). To improve real-time recognition efficiency of martial arts movements, Mediapipe is used to extract standardized human joint sequences for accurate 3D pose estimation. The process is shown in Fig. 1.

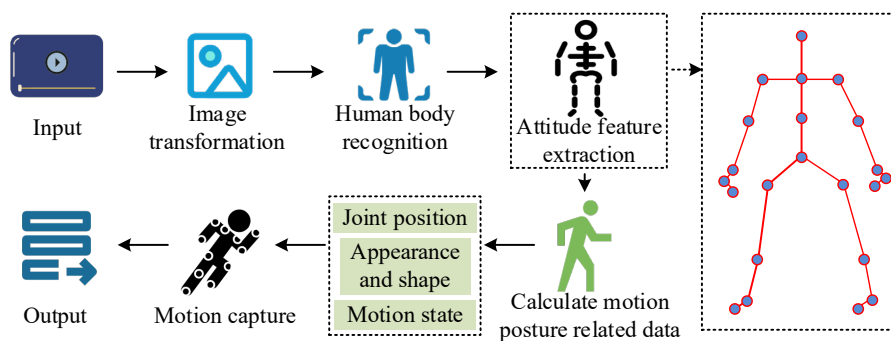


Fig. 1. Human joint feature extraction using Mediapipe

As shown in Fig. 1, this study uses Mediapipe to perform frame-by-frame detection of Wushu actions in videos. Image transformation normalizes the pixel coordinates of joint nodes and determines the spatial positions and scales of human actions. From the perspective of action recognition theory, the semantics of human actions are not solely determined by a single frame of static posture, but are characterized by the spatial configuration of static posture and the temporal evolution of dynamic trajectory. Static posture reflects the instantaneous structural characteristics of a specific action, while dynamic trajectory characterizes the motion pattern of limbs on a continuous time axis. The two aspects complement each other and are indispensable, forming a dual foundation for understanding actions. Therefore, based on the above theory, the method extracts joint node features from human pose data. The extracted joints include key posture feature points of the limbs and the body. By analyzing and predicting human keypoints, the method computes feature angles and motion trajectory

parameters. It captures the posture and position of Wushu actions. Vector geometry methods calculate Wushu pose information, including joint positions of key limbs, appearance shape, and motion velocity. The mathematical expression for joint node parameterization is given by Eq. (1).

$$X_i = \sum_{j=1}^I \omega_{i,j} T_j(\theta, C) T_j(\bar{\theta}, C)^{-1} [\tilde{X}_i + \Delta \tilde{X}_i, \Delta \tilde{X}_i^f] \quad (1)$$

In Eq. (1), j denotes joint nodes, $T_j(\theta, C)$ denotes the transformation matrix, I denotes the number of nodes, and $\bar{\theta}$ denotes the static posture. Although the Mediapipe-based method can extract accurate human joint coordinates, it cannot capture the intrinsic temporal correlation among skeleton sequences. In contrast, the Transformer algorithm enhances the vector representation of data, mitigates vanishing or exploding gradients, and efficiently establishes weight associations (Papa et al., 2024). Therefore, to capture dynamic associations and long-term dependencies in martial arts action sequences, the study introduces the Transformer algorithm and develops the Mediapipe Transformer algorithm, thereby enhancing its performance and generalization. The specific mechanism of the Mediapipe Transformer algorithm for recognizing features of martial arts actions is shown in Fig. 2.

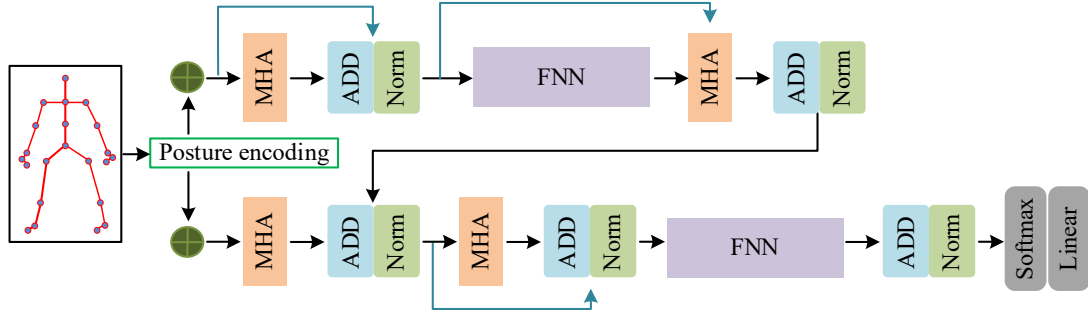


Fig. 2. Mediapipe-Transformer mechanism for Wushu action recognition

As shown in Fig. 2, Mediapipe-Transformer determines the position and scale of the human action through frame-by-frame detection and extracts joint node features via keypoint detection. The method then inputs joint nodes into the Transformer and processes the information using an encoder-decoder architecture. The encoder and decoder consist of MHA and FNN. Each layer introduces residual connections and layer normalization. MHA captures Wushu pose information from different perspectives. The mathematical formulation of the attention mechanism is shown in Eq. (2).

$$Attention(Q, K, V) = \text{soft max}(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

MHA performs self-attention operations independently in each subspace. Its calculation is shown in Eq. (3).

$$MultiHead(Q, K, V) = \text{concat}(head_1, head_2, \dots, head_k) \quad (3)$$

In Eq. (3), *concat* denotes the relationship modeling process. The output of MHA is fed into FNN for feature transformation. Finally, the high-dimensional feature vector of martial arts posture is obtained through the layer normalization module. Although Mediapipe-based pose recognition can obtain accurate joint coordinates, it has limitations in modeling temporal dependencies. Transformer enhances vector representation, alleviates gradient issues, and efficiently establishes feature associations. To this end, the study introduces a Transformer-based algorithm, Mediapipe Transformer, which captures dynamic correlations and long-term dependencies in martial arts movements, thereby improving model performance and generalization. The mechanism is shown in Fig. 3.

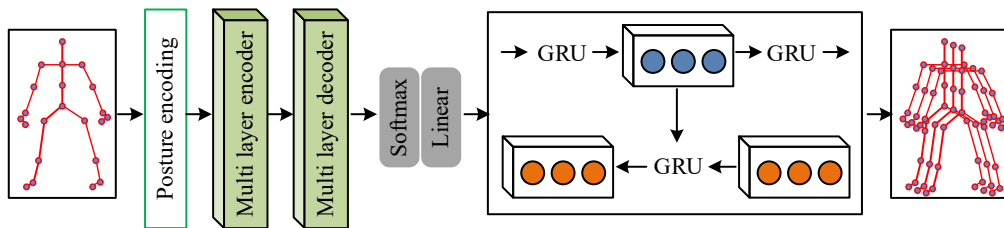


Fig. 3. Mediapipe-Transformer-BiGRU operation flow

As shown in Fig. 3, Mediapipe first processes the original martial arts action video sequence to extract joint key points. Subsequently, a transformer is used to capture spatiotemporal order. Nonlinear transformations are achieved through multi-layer encoding and decoding. Finally, BiGRU processes the skeleton sequence forward in time through the forward GRU

layer, simulating the natural sequence of martial arts 3D poses and gradually accumulating and transmitting historical information. The backward GRU layer processes the sequence in reverse, capturing future contextual information along the timeline. BiGRU concatenates the outputs of the forward and backward layers to form fused features. The update gate determines how many memories are retained in the current state, thereby controlling the continuity of action features. The specific calculation expression is shown in Eq. (4).

$$z_t = \sigma(W \cdot x_t + W \cdot h_{t-1} + b_z) \quad (4)$$

In Eq. (4), W denotes the parameter matrix, h_{t-1} denotes neuron output, and σ denotes the Sigmoid function. The reset gate computes a new Wushu 3D pose candidate state and enables the forgetting of local information. Its mathematical expression is shown in Eq. (5).

$$r_t = \sigma(W \cdot x_t + W \cdot h_{t-1} + b_r) \quad (5)$$

The method combines the current input with the reset-gated past state to generate a Wushu 3D pose candidate state. The calculation is shown in Eq. (6).

$$\tilde{h}_t = \tanh(W \cdot x_t + W \cdot (r_t \Theta h_{t-1}) + b_{\tilde{h}}) \quad (6)$$

In Eq. (6), Θ denotes the Hadamard product. The update gate fuses the previous state with the new candidate to produce the hidden state at the current time step. The calculation is shown in Eq. (7).

$$h_t = (1 - z_t) \Theta \tilde{h}_t + z_t \Theta h_{t-1} \quad (7)$$

2.2. Construction of Wushu 3D Pose Recognition Model Based on DSC-MGTB

Although the Mediapipe Transformer BiGRU algorithm can effectively establish dependency associations and local continuity of actions, it has limitations in capturing the inherent spatial structure of human joints. GCN can directly model the human skeleton as a graph structure and learn the inherent spatial connections and mechanical relationships between joints (Zhang et al., 2023). Therefore, to enhance the algorithm's ability to capture the inherent spatial structure of human joints, the Mediapipe Transformer BiGRU and GCN algorithms are fused, and the resulting algorithm is named MGTB to comprehensively improve deep feature representation and recognition accuracy for martial arts movements. The running process of the MGTB algorithm is shown in Fig. 4.

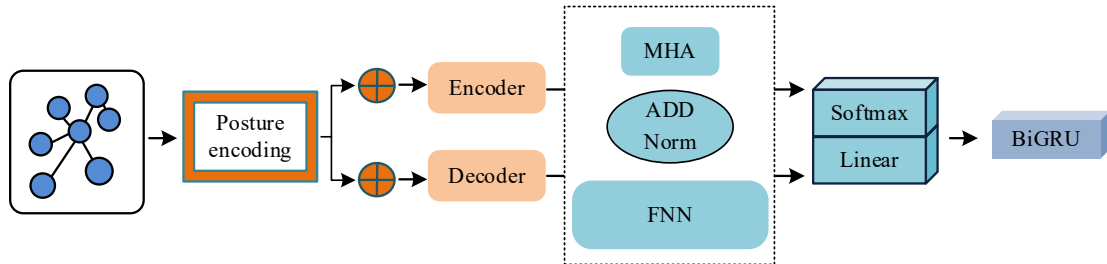


Fig. 4 MGTB algorithm operation flow

As shown in Fig. 4, the MGTB algorithm first uses Mediapipe to extract key points of the human body structure, and inputs the constructed human skeleton map into GCN for processing. The GCN algorithm takes the human skeleton structure as input, convolves over joint-node data, and integrates adjacent nodes. Subsequently, the node features are iteratively updated in the hidden layer based on the convolution results, and the receptive field is expanded by stacking the number of network layers. After aggregating feature nodes, the contextual features are fused with the nodes' original motion features to enrich each node's information. The study serializes the output data features into data blocks, with the calculation formula for the serialized features given by Eq. (8).

$$K = \text{Reshape}\left(\frac{N}{R}, C \cdot R\right)k \quad (8)$$

In Eq. (8), Reshape denotes serialization, N denotes the number of channels, R denotes the reduction ratio, C denotes reshaped features, and k denotes original features. The serialized sequence is further reduced in dimension. The calculation is shown in Eq. (9).

$$K_1 = \text{Linear}(C \cdot R, C)K \quad (9)$$

In Eq. (9), Linear denotes linear processing. Subsequently, the joint correlation weights are modeled with a Transformer, and the temporal dependencies among actions are captured with a BiGRU. Although MGTB can capture complex spatiotemporal features, it is computationally expensive. DSC can significantly reduce parameter count and computational complexity through deep and pointwise convolutions. Therefore, to improve the computational and storage

efficiency of the model, DSC is introduced into MGTB to construct a DSC-MGTB martial arts 3D pose recognition model, thereby compressing and reducing the dimensionality of the intermediate feature maps within the network. The operation process of the DSC-MGTB model is shown in Fig. 5.

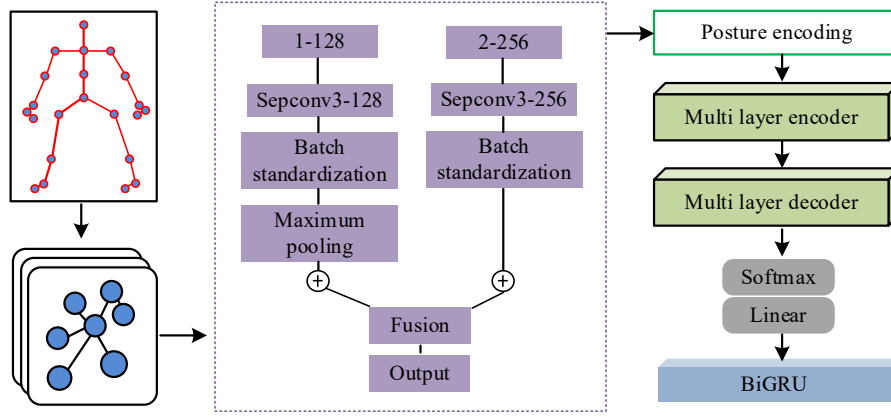


Fig. 5. DSC-MGTB model operation flow

As shown in Fig. 5, the model constructs a static human skeleton through Mediapipe and uses GCN to adaptively learn joint weights. DSC then performs multi-scale convolution on the data. Fusion of feature maps with different sizes achieves compression and dimensionality reduction. The convolution operation is shown in Eq. (10).

$$x_i^l = f(W_i^l * X^{(l-1)} + b_i^l) \quad (10)$$

In Eq. (10), x_i^l denotes the i -th feature of the l layer, and W_i^l denotes the weight matrix. Subsequently, spatial correlation is modeled using deep and pointwise convolutions to enhance feature representation. Small-scale convolutional kernels capture local motion details, while large-scale convolutional kernels perceive overall limb motion trends. Multi-scale convolution fusion can reduce the size and number of channels in feature maps, and its formula is given by Eq. (11).

$$M_s = \sigma(\text{conv}(F_{\text{concat}})) \quad (11)$$

In Eq. (11), M_s denotes convolution kernels, and F_{concat} denotes feature map concatenation. The Transformer then establishes global spatiotemporal dependencies, and the BiGRU outputs 3D Wushu action feature vectors to achieve 3D Wushu pose recognition.

3. Experimental Results

For the experiment, Windows 11 and Ubuntu 18.04.01 were chosen as the operating systems, Intel(R) Xeon(R) Platinum as the processor, and PyTorch 1.8 as the deep learning framework. The specific experimental parameters are shown in Table 1.

The dataset selected is the Martial Arts, Dance and Sports (MADS) dataset, released by City University of Hong Kong in 2017, specifically designed for 3D human pose estimation and action recognition tasks. This dataset contains approximately 53,000 image frames (24 GB), with 2 martial arts masters, 2 dancers, and 1 athlete as participants. The movements cover 5 categories, including Tai Chi, Karate, Jazz Dance, Hip Hop Dance, and sports combinations, with 6 sequences for each category. Three cameras capture data at 15 fps with a resolution of 1024×768 , while the motion capture system operates at 60 fps. This study selected two types of images, Tai Chi and Karate, totaling 21200 frames, and divided them into training and testing sets in a 7:3 ratio.

3.1. Performance Validation of MGTB

To evaluate the performance of the MGTB algorithm in Wushu 3D pose recognition tasks, this study conducted comparative experiments with Multiscale Vision Transformers (MVit), Dynamic Multi-scale Graph Neural Network (DMGNN), and Bidirectional Long Short-Term Memory Network (BiLSTM). To analyze the accuracy and loss of the MGTB algorithm in Wushu 3D pose recognition, this study conducted comparative tests, with results shown in Fig. 6.

As shown in Fig. 6(a), the accuracy of the MGTB algorithm increased rapidly to 95.18% in the first 10 iterations, then converged smoothly to 97.01% after 100 iterations. Fig. 6(b) shows that the loss of the MGTB algorithm decreased from 0.27 in the initial iteration to 0.04 in the final iteration. Overall, the proposed algorithm outperformed the other three algorithms in terms of accuracy and loss, indicating that it achieves higher recognition accuracy and stability. To further analyze the operational performance of the MGTB algorithm in Wushu 3D pose recognition tasks, this study compared the fitness and response time of the four algorithms, with results shown in Fig. 7.

As shown in Fig. 7(a), the fitness curve of the proposed algorithm dropped sharply in the first 20 iterations, then flattened, reaching an optimal value of 5.13, which is superior to those of the other three algorithms. Fig. 7(b) indicates that the algorithm's response time grows slowly with iterations, reaching only 109.8ms at 100 iterations, showing better performance. The results showed that MGTB converged and responded quickly. To further assess the necessity of the

MGTB hybrid algorithm, an ablation study was conducted, including Algorithms A (full MGTB), B (A without GCN), C (B without the BiGRU bidirectional structure), and D (C without the entire BiGRU). Macro F1 scores and recall rates for martial arts pose recognition were evaluated, with results shown in Fig. 8.

Table 1. Experimental parameter settings

Parameter Category	Specific Configuration	MGTB	DSC-MGTB
Hyperparameters	Batch Size	32	32
	Initial Learning Rate	0.001	0.001
	Optimizer	Adam	Adam
	Training Epochs	100	100
	Sequence Length	64 frames	64 frames
	Number of GCN Layers	3	3
	Number of Transformer Layers	4	4
	Number of Transformer Attention Heads	8	8
Model Scale	Number of BiGRU Hidden Units	256	128
	Total Parameters	8.2 M	3.1 M
	FLOPs per Sample	2.1 G	0.8 G
Training Time	Time per Epoch	85 seconds	35 seconds
	Total Training Time (100 Epochs)	2.4 hours	1.0 hour
Computational Cost	Peak VRAM Usage	6.2 GB	2.8 GB
	Inference Speed (FPS)	120 FPS	280 FPS

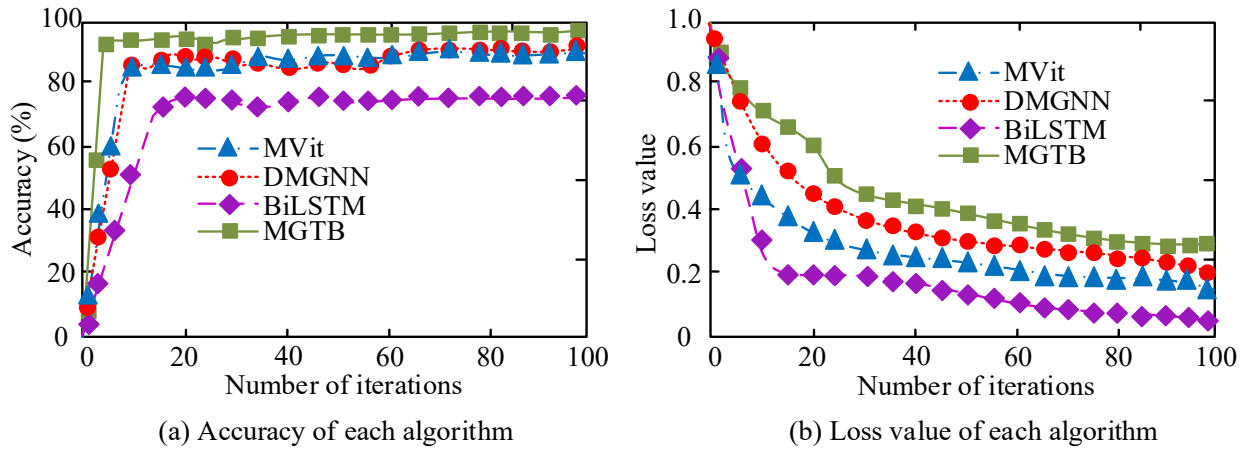


Fig. 6. Accuracy and loss comparison of different algorithms

As shown in Fig. 8(a), the median macro F1 score of the complete Algorithm A was 92.74%. As modules were removed, the algorithms' macro F1 score decreased accordingly, with Algorithm D's dropping to 81.13%. As shown in Fig. 8(b), the median recall rate of the complete Algorithm A was 94.52%. After removing the GCN module, Algorithm B's median recall decreased to 87.08%. After removing the bidirectional structure of the BiGRU module, the median recall rate of Algorithm C was 83.54%, and finally, Algorithm D's recall rate decreased to 78.14%. The results demonstrated the effectiveness and necessity of the complete algorithm structure proposed in this study.

3.2. Analysis of Pose Recognition Performance of DSC-MGTB Model

After validating the MGTB algorithm's performance in martial arts 3D pose recognition, further analysis of the DSC-MGTB model was conducted by comparing it with 3D pose recognition models built on MVit, DMGNN, and BiLSTM algorithms. The experiment used an Intel (R) Xeon (R) Gold 6346 CPU, an RTX 4090 GPU, and 48GB of RAM, with a learning rate of 0.0001 and up to 100 iterations. The MADS dataset was selected, with keyframes uniformly extracted from Tai Chi and Wing Chun videos. To assess the DSC-MGTB model's 3D pose recognition, we compared the motion capture trajectories of the four models, as shown in Fig. 9.

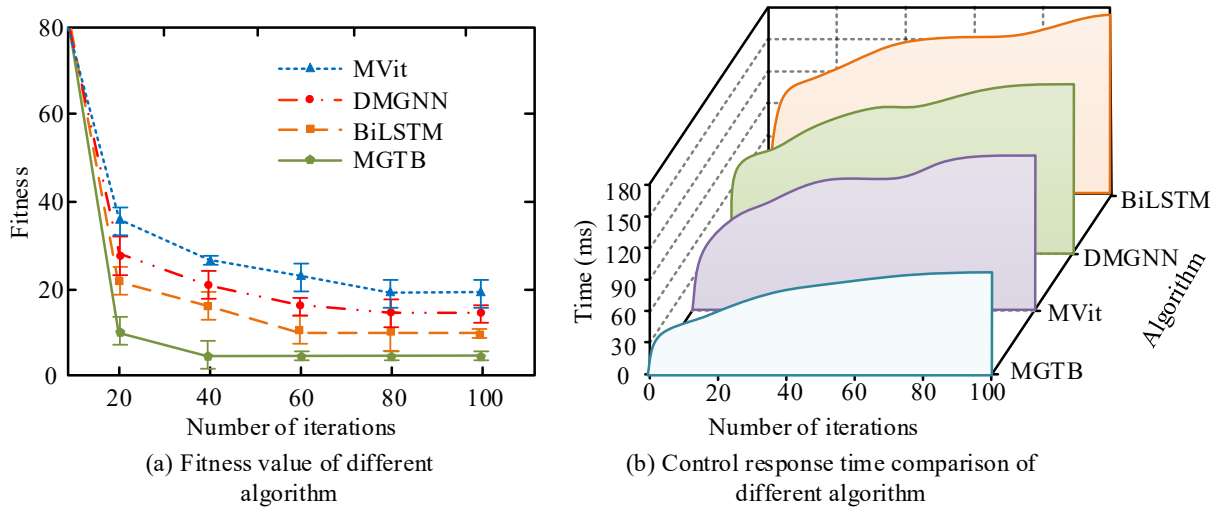


Fig. 7. Fitness and response time comparison of different algorithms

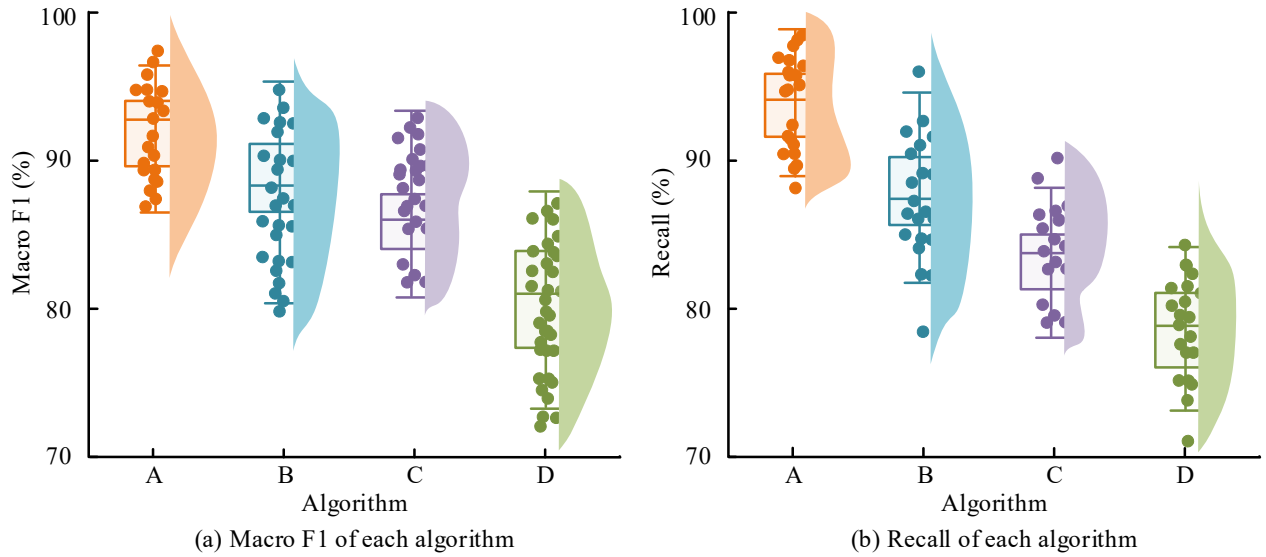


Fig. 8. Ablation experiment results

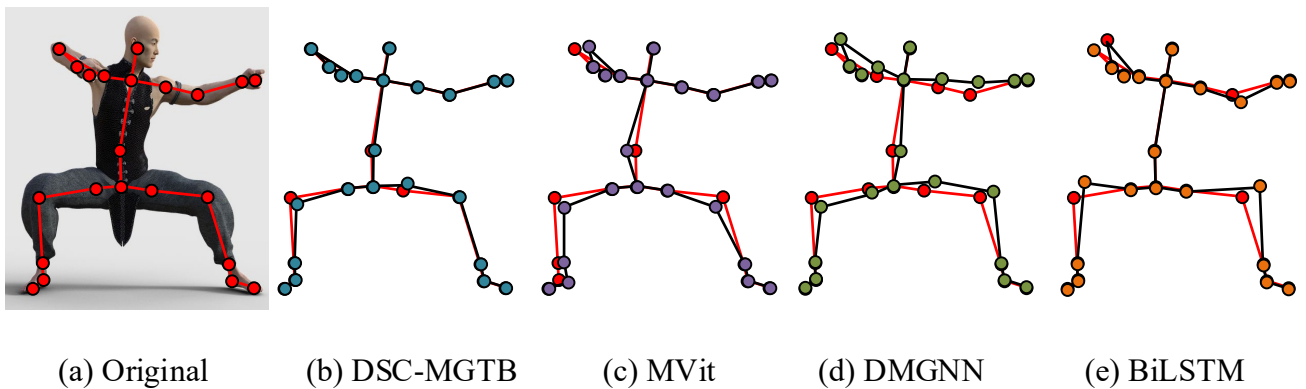


Fig. 9. Wushu action capture results comparison (source: <https://pixabay.com/illustrations/kung-fu-martial-arts-pose-fighter-4938609/>)

As illustrated in Fig. 9(b), the martial arts posture trajectory identified by the DSC-MGTB model demonstrates a high degree of overlap with the actual action posture depicted in Fig. 9(a) at key nodes. As shown in Fig. 9(c), (d), and (e), the MVit model exhibits offsets in the key nodes of the right elbow and lower limbs, the DMGNN model shows varying degrees of deviation in the key nodes of the limbs and body, and the BiLSTM model demonstrates significant deviations in all limbs. In summary, the DSC-MGTB model outperforms the comparative models in accuracy for martial arts action

recognition. To analyze the Wushu movement capture accuracy of the DSC-MGTB model, this study varied the number of people and conducted comparative tests on the Mean Per Joint Position Error (MPJPE) and Percentage of Correct Keypoints (PCK) across the four models, with the results shown in Fig. 10.

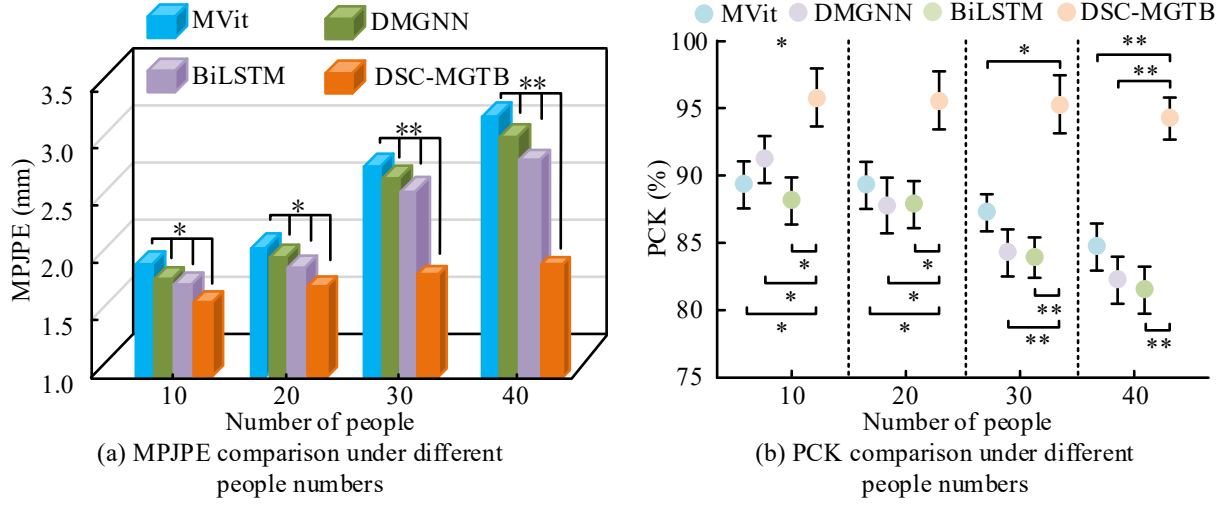


Fig. 10. MPJPE and PCK comparison of different models

Note: * indicates $p < 0.05$ compared to the comparative model, ** indicates $p < 0.01$, which is statistically significant.

As shown in Fig. 10(a), the MPJPE of DSC-MGTB was 1.66 mm when the number of people was 10, and 1.99mm when the number of people increased to 40, which was significantly better than the comparison models in all scenarios ($p < 0.01$). As shown in Fig. 10(b), when the number of people is 10, the average PCK of the DSC-MGTB model is 96.63%, which is higher than the control model ($p < 0.05$). As the number of people increases, the PCK of the DSC-MGTB model ultimately reaches 94.32%, which is significantly higher than that of the control model ($p < 0.01$). In summary, the proposed model demonstrated high accuracy in recognizing 3D Wushu movements. To further analyze the 3D Wushu movement recognition precision of the DSC-MGTB model, this study evaluated the mean Average Precision (mAP) of the four models at Intersection over Union (IoU) thresholds of 0.5 and 0.95, with results shown in Fig. 11.

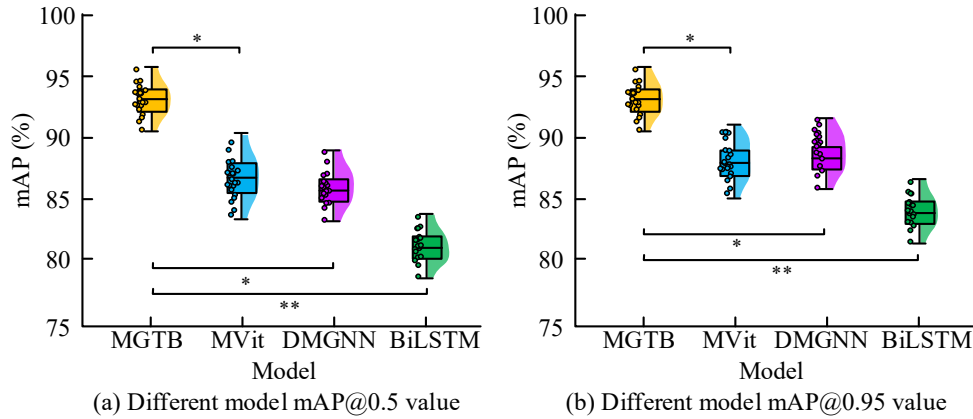


Fig. 11. mAP comparison under different interaction ratios

As shown in Fig. 11(a), with an IoU threshold of 0.5, the median mAP for action recognition in the DSC-MGTB model reached 93.54%, while the mAP values of the other three models were lower than those of the proposed model ($p < 0.05$). As shown in Fig. 11(b), with an interaction ratio of 0.95, the median mAP for action recognition in the DSC-MGTB model reached 93.42%, significantly higher than the control model ($p < 0.05$). The results above indicate that the DSC-MGTB model achieves superior accuracy for martial arts action recognition. To further explore the 3D Wushu pose recognition performance of the DSC-MGTB model, this study compared the recognition accuracy, recall rate, and F1 score of the four models in two Wushu scenarios: Tai Chi and Wing Chun, with results shown in Table 2.

As shown in Table 2, in the Tai Chi scene, the DSC-MGTB model achieves recognition accuracy of 95.96%, recall of 96.27%, and F1 score of 96.24%, all of which are higher than those of the comparison models. In the Wing Chun scene, the DSC-MGTB model has a posture recognition accuracy of 95.87%, which was 13.46%, 9.28%, and 13.03% higher than those of the comparison models, respectively. At the same time, the recall of the research model and F1 were 96.99% and 96.34%, respectively, significantly higher than those of the comparison model. Overall, the DSC-MGTB model has high accuracy in martial arts posture recognition. After verifying the recognition accuracy and computational efficiency of the DSC-MGTB model, further research was conducted to explore its practical application value in martial arts training

management scenarios. The model score is calculated by combining MPJPE weighting, a joint-angle-over-limit penalty term, and dynamic time regularization similarity, with a score range of 0-100. The quantitative evaluation results of the DSC-MGTB model for typical martial arts movements, including key joint deviations, comprehensive scores, and feedback suggestions, are shown in Table 3.

Table 2. Wushu pose recognition performance of different models

Scenario	Model	Accuracy (%)	Recall (%)	F1 score (%)
Tai Chi	MVit	85.43	83.25	84.12
	DMGNN	81.42	85.39	80.23
	BiLSTM	83.30	87.45	88.88
	DSC-MGTB	95.96	96.27	96.24
Wing Chun	MVit	82.41	89.36	97.39
	DMGNN	86.59	84.71	85.04
	BiLSTM	82.84	81.28	81.55
	DSC-MGTB	95.87	96.99	96.34

Table 3. Quantitative evaluation for martial arts movements

Movement name	Trainee ID	Key joint deviations	Model score	Feedback suggestions
Tai Chi-Cloud Hands	101	Left elbow angle deviation of +8.5°, right wrist trajectory offset of 6.2%	82.3	Sink left elbow; reduce right hand arc
Wing Chun-Sun-style Punch	205	Right shoulder-elbow force chain timing delay of 0.12s, insufficient fist surface rotation	76.8	Elbow-driven force; rotate fist during punch
Wing Chun-Tan Da	308	Tan hand angle deviation of +6.3°, punching path offset	68.5	Simultaneous defense/counterattack; midline focus
Tai Ch-Brush Knee and Twist Step	112	Excessive center of gravity height fluctuation, knee inward inclination	88.1	Maintain stable posture; align knees with toes

Table 3 presents the quantitative evaluation and feedback results of the DSC-MGTB model across four sets of martial arts movements. The model can automatically identify and quantify key joint deviations, including a +8.5 ° deviation in the left elbow angle of Tai Chi - Yun Hand and a 0.12 s delay in the force chain timing in Wing Chun - Ri Zi Chong Quan. At the same time, the model provides a comprehensive score on a percentage scale. In addition, the text feedback suggestions generated by the model have clear operational directionality and directly correspond to common training errors. Based on the model's output, managers can implement differentiated decisions. Overall, the DSC-MGTB model has practical value for real-world martial arts training and management scenarios.

4. Interpretation of Results

The proposed MGTB algorithm showed significant advantages in 3D martial arts pose recognition. Experiments achieve 97.01% accuracy and a loss of 0.04, reducing loss while maintaining high accuracy. This stems from its layer-by-layer extraction and fusion of deep features, enabling faster convergence to an optimal loss surface. Zhou et al. (2023) proposed a sports action model based on adaptive binary particle swarm optimization that accurately extracted features and exhibited a similar performance trend. MGTB also achieves optimal fitness swiftly, reducing response time. Its fast convergence and efficiency stem from using a Transformer for long-range dependencies and a BiGRU to refine temporal dynamics, thereby reducing the number of iterations and response time. Shi et al. (2023) used a multi-stream architecture and content-adaptive attention to focus on key body parts for accurate recognition. Compared with the method proposed by Shi et al. (2023), MGTB converged faster.

During verification of the DSC-MGTB model's 3D martial arts pose recognition, it accurately detected joint changes, achieving an MPJPE of 1.99 mm and a PCK of up to 94.32%. The model's balance of joint-motion spatial details enhances recognition accuracy, aligning with the findings of Wu and Chen (2023). Meanwhile, the DSC-MGTB model achieved recognition accuracies of 95.96% and 95.87% across different scenarios. These results surpassed those of the comparison models because DSC reduced model complexity while maintaining high feature-extraction efficiency and discriminative power with limited resources. Zhao et al. (2023) improved recognition robustness under occlusion and interference by fusing multimodal sensor data with spatiotemporal graph convolutional networks, reaching 92.8% accuracy. Unlike their approach, the DSC-MGTB model achieved higher accuracy by efficiently fusing internal algorithms. In addition, the DSC-MGTB model can accurately quantify the evaluation and feedback output of martial arts movements, providing guidance

for managers to implement differentiated decisions. Managers can promote specialized training videos and arrange decomposed exercises for students with low scores and significant deviations. Students with common errors can be grouped into targeted reinforcement training sessions. For students with standardized movements, they can be used as teaching demonstrations in recorded videos. Overall, the research model effectively transformed technical indicators into actionable management strategies.

5. Conclusion

In response to the limitations of low efficiency and poor adaptability in traditional 3D pose recognition technology, this study extracts joint points using Mediapipe and integrates Transformer and BiGRU algorithms to capture global spatiotemporal dependencies, thereby enabling pose extraction and enhancement. GCN is introduced to simultaneously highlight the spatial structure and correlations among core joints in martial arts movements. On this basis, a martial arts 3D pose recognition model is constructed by combining DSC. The results indicate that the MGTB algorithm has a fast response and high accuracy. The DSC-MGTB model has superior recognition accuracy and lower joint positioning error. The proposed model demonstrated high accuracy across different IoU thresholds and scenarios. In addition, the proposed model can quantitatively evaluate different martial arts movements and provide feedback. Overall, the proposed model demonstrated superior accuracy in martial arts 3D pose recognition, reduced joint positioning errors, and maintained operational stability.

Although the study's data collection focuses on two typical martial arts, Tai Chi and Wing Chun, the proposed technical framework demonstrates strong generalization and cross-cultural adaptability. Mediapipe, Transformer, GCN, and BiGRU are all data-driven universal modules that do not rely on the physical characteristics or regional culture of specific martial arts schools. At the same time, this model only requires monocular RGB video input and does not require high-cost motion capture equipment or specific venues. It can be easily applied to martial arts training centers, school physical education classrooms, and online teaching platforms worldwide. Therefore, the proposed technical solution has the potential to be adopted across various martial arts systems worldwide, providing universal technical support for digital teaching and standardized martial arts evaluation.

Managers of martial arts training institutions, sports schools, and teams can optimize decision-making based on this approach: (1) Introducing models to obtain quantitative scores of student movements and joint deviation reports, improving feedback objectivity and efficiency. (2) Utilizing model intelligence to identify errors and concentrating human resources on individualized correction and tactical guidance. (3) Automatically evaluate videos uploaded by remote platform students and establish a unified action specification level certification system. (4) The model supports both regular cameras and smartphones, reducing hardware barriers and budget.

This study demonstrated an efficient martial arts posture recognition algorithm and illustrated the application mechanism of a deep learning architecture in human motion analysis. The research approach for the perception of complex movements by models can be further extended to a broader field of human motion science, such as deploying the model on AR glasses or mobile apps to achieve motion guidance and correction in sports fields such as martial arts, dance choreography, rehabilitation training, and even human-computer interaction. In addition, this study did not validate generalization performance under severe occlusion and variable lighting conditions. Therefore, future research can examine the stability of the model's pose estimation under complex conditions, such as varying lighting and repeated occlusions.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author confirms that no AI tools were used in the preparation of this manuscript.

References

- Ashtari-Majlan, M. and Masip, D. (2025). Spatial-aware transformer-GRU framework for enhanced glaucoma diagnosis from 3D OCT imaging. *IEEE Journal of Biomedical and Health Informatics*, 29(9), 6167-6175.
- Chauhan, N. S. and Kumar, N. (2024). Novel confined attention-enabled hierarchical Bi-LSTM and Bi-GRU fusion: a multiscale traffic flow prediction application. *IEEE Transactions on Computational Social Systems*, 11(6), 7541-7554.
- Chen, Q. (2025). Stock price prediction with LLM-guided market movement signals and transformer model. *FinTech and Sustainable Innovation*, 1, A12-A12.
- Guo, X., Zhu, Q., Wang, Y., and Mo, Y. (2024). MG-GCT: A motion-guided graph convolutional transformer for traffic gesture recognition. *IEEE Transactions on Intelligent Transportation Systems*, 25(10), 14031-14039.
- Helmy, I., Tarafder, P., and Choi, W. (2023). LSTM-GRU model-based channel prediction for one-bit massive MIMO system. *IEEE Transactions on Vehicular Technology*, 72(8), 11053-11057.
- Kishore, P. V. V. and Kumar, A. (2025). Est3D2Real-estimated 3D-to-real data embeddings for real time sign language recognizer. *Pattern Recognition Letters*, 196(1), 86-92.
- Li, Y. and Wu, L. (2023). A wireless self-powered sensor network based on dual-model convolutional neural network algorithm for tennis sports. *IEEE Sensors Journal*, 23(18), 20745-20755.
- Mokayed, H., Quan, T. Z., Alkhaled, L., and Sivakumar, V. (2023). Real-time human detection and counting system using deep learning computer vision techniques. *Artificial Intelligence Applications*, 1(4), 221-229.

- Myung, W., Su, N., Xue, J. H., and Wang, G. (2024). Degen: Deformable graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 33, 2477-2490.
- Papa, L., Russo, P., Amerini, I., and Zhou, L. (2024). A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 7682-7700.
- Pradani, K. A. and Suadaa, L. H. (2023). Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 10(6), 1177-1184.
- Roccliffe, P., Adamakis, M., O'Keeffe, B. T., Walsh, L., Bannon, A., Garcia-Gonzalez, L., Chambers, F., Stylianou, M., Sherwin, I., Mannix-McNamara, P., and MacDonncha, C. (2024). The impact of typical school provision of physical education, physical activity and sports on adolescent mental health and wellbeing: A systematic literature review. *Adolescent Research Review*, 9(2), 339-364.
- Sánchez-Brizuela, G., Cisnal, A., de la Fuente-Lopez, E., Fraile, J. C., and Perez-Turiel, J. (2023). Lightweight real-time hand segmentation leveraging MediaPipe landmark detection. *Virtual Reality*, 27(4), 3125-3132.
- Shi, W., Li, D., Wen, Y., and Yang, W. (2023). Occlusion-aware graph neural networks for skeleton action recognition. *IEEE Transactions on Industrial Informatics*, 19(10), 10288-10298.
- Vu, C. C., Nguyen, T. L., and Nguyen, T. H. (2024). Finger gesture recognition by MediaPipe algorithm and advanced YOLOv7 network for deaf people. *International Journal of Biomedical Engineering and Technology*, 48(1), 55-74.
- Wu, F. and Chen, H. (2023). The analysis of technical characteristics of badminton for sports with neurorobotics under machine learning. *IEEE Access*, 11, 144337-144348.
- Xu, M., Guo, L., and Wu, H. C. (2023). Robust abnormal human-posture recognition using OpenPose and multiview cross-information. *IEEE Sensors Journal*, 23(11), 12370-12379.
- Zhang, J., Gong, K., Wang, X., and Feng, J. (2023). Learning to augment poses for 3D human pose estimation in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8), 10012-10026.
- Zhang, J., Li, X., Wang, Y., Wang, C., Yang, Y., Liu, Y., and Tao, D. (2024). Eatformer: Improving vision transformer inspired by evolutionary algorithm. *International Journal of Computer Vision*, 132(9), 3509-3536.
- Zhao, C., Wang, L., Xiong, F., Chen, S., Su, J., and Xu, H. (2023). RFID-based human action recognition through spatiotemporal graph convolutional neural network. *IEEE Internet of Things Journal*, 10(22), 19898-19912.
- Zhou, Y., Wang, R., Wang, Y., Sun, S., Chen, J., and Zhang, X. (2023). A swarm intelligence assisted iot-based activity recognition system for basketball rookies. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1), 82-94.



Yufeng Ma was born in 1977 in Weifang, Shandong, China. He earned his master's degree from Shanghai University of Sport in 2006, with a major in traditional ethnic sports. Currently serving as an associate professor at Nanjing Police University. His research specializes in Wu Shu training theory and methodology, with a particular emphasis on practical instruction and training for police operations.