IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

**Traitement du Signal**
Vol. 43, No. 3, June, 2026, pp. 1507-1522
Journal homepage: http://iieta.org/journals/ts

# A Deep Learning-Based Real-Time Action Recognition and Analysis System for Martial Arts Athletes

Tianyi Zhang

School of Physical Education, Shandong Normal University, Jinan 250358, China

Corresponding Author Email: 202415210116@stu.sdnu.edu.cn



https://doi.org/10.18280/ts.430333

**Received:** 2 April 2026
**Revised:** 29 May 2026
**Accepted:** 9 June 2026
**Available online:** 30 June 2026

***Keywords:***
*martial arts action recognition, graph convolutional networks, spatiotemporal feature modeling, lightweight model, multimodal fusion, intelligent sports*

**ABSTRACT**

Martial arts movements are characterized by strong dynamic joint coordination, pronounced phased motion rhythms, and wide ranges of limb motion amplitude. Conventional graph convolutional networks are inadequate for modeling the complex spatiotemporal dynamics inherent in martial arts and hinder their applicability in intelligent training and real-time motion analysis for martial arts. To address these issues, a deep learning system tailored for real-time martial arts athlete action recognition and analysis was developed. First, a deformable adaptive graph convolution module was proposed, integrating anatomical priors with a data-driven feature association learning mechanism to dynamically adjust the receptive field of joint neighborhoods, thereby precisely modeling the evolving joint coordination patterns in martial arts movements. Second, a martial-arts-semantics-driven multi-scale temporal modeling and phase perception module was established. Unsupervised temporal clustering was employed for automatic action phase segmentation, and temporal convolution scales were adaptively matched according to the rhythmic characteristics of distinct movement phases. An attention mechanism was further incorporated to enhance feature weights during critical force-generating phases, enabling comprehensive extraction of phase-specific temporal semantic features. Simultaneously, model lightweighting was achieved through the integration of grouped graph convolution, temporal sparse sampling, and knowledge distillation strategies. A cross-attention fusion mechanism between Red-Green-Blue (RGB) and skeletal modalities was constructed to compensate for unimodal feature deficiencies in complex scenes and to enhance environmental robustness. This study effectively overcomes the modeling bottlenecks of conventional skeleton-based action recognition models in specialized martial arts scenarios, achieving high-accuracy, low-latency intelligent martial arts movement recognition. The proposed system provides reliable technical support for quantitative movement analysis, scientifically informed training guidance, and competitive performance evaluation in martial arts.

## 1. INTRODUCTION

Human action recognition constitutes a core research direction within the field of computer vision for intelligent perception, possessing broad application prospects in areas such as human-computer interaction, intelligent surveillance, and quantitative motion analysis [1, 2]. In comparison with conventional visible-light images and depth images, sequences of human skeletal keypoints effectively circumvent interference caused by illumination fluctuations, background clutter, and variations in viewing angles, while precisely characterizing the intrinsic laws of human limb motion, thereby serving as a high-quality data modality for high-precision human action modeling [3, 4]. Graph convolutional networks, by virtue of their modeling advantages in adapting to irregular skeletal topological structures, have transcended the limitations of conventional convolutional networks in structured human motion modeling, and have now become the prevailing technical paradigm for skeleton-based human action recognition, continuously driving performance improvements in general action recognition tasks [5, 6]. As a typical high-complexity specialized sport, martial arts movements are characterized by both variable joint coordination mechanisms and hierarchical rhythmic features, with large limb motion amplitudes, pronounced instantaneous deformations, and clearly distinguishable force-generating phases, imposing technical demands on dynamic adaptability and fine-grained representational capacity in spatiotemporal feature modeling that far exceed those of conventional daily actions [7, 8]. Current martial arts training and competitive evaluation are progressively transitioning from experience-based guidance toward digital, intelligent, and quantitative analysis paradigms, necessitating real-time action recognition technologies that combine high accuracy with low latency for the purposes of athlete movement standardization verification, quantitative analysis of force-generation characteristics, and intelligent optimization of training regimens [9, 10]. Consequently, the investigation of a specialized real-time action recognition and analysis system for martial arts not only serves to advance the application framework of graph-based

spatiotemporal modeling theory in complex dynamic motion scenarios and fill the technical gap in specialized martial arts intelligent recognition, but also provides robust technical support for scientifically grounded martial arts training, competitive performance assessment, and intelligent dissemination of sports culture, thereby holding significant academic value and engineering application value.

At present, skeleton-based action recognition techniques employing graph convolutional networks have achieved remarkable recognition performance on general-purpose action datasets. However, these general-purpose models exhibit significant technical deficiencies when adapted to the complex motion scenarios characteristic of specialized martial arts, rendering them inadequate for meeting the application demands of high-precision real-time recognition [11, 12]. At the level of spatial topology modeling, conventional graph convolutional networks rely on fixed human anatomical structures to construct adjacency relationships for feature propagation, with the topological structure remaining invariant throughout the feature learning process, thereby failing to accommodate the dynamically changing joint coordination associations under the substantial limb deformations inherent in martial arts movements [13, 14]. Although existing adaptive graph convolution methods can learn joint weight associations through data-driven approaches, a globally uniform weight allocation strategy is predominantly adopted, which precludes the dynamic adjustment of node receptive field ranges according to instantaneous motion amplitudes and limb states, thus rendering the capture of transient and variable joint coupling characteristics in martial arts movements difficult, and resulting in severely insufficient dynamic adaptability in spatial modeling [15, 16]. At the level of temporal feature modeling, existing temporal modeling methods generally employ fixed-scale convolutional kernels to uniformly process complete video sequences, while neglecting the inherently phased motion characteristics of martial arts movements [17, 18]. A complete martial arts movement sequence encompasses progressive motion phases, with significant differences in motion speed, limb amplitude, and force-generation logic across distinct phases. The homogenized temporal feature extraction approach fails to differentiate the feature importance of each motion phase, thereby attenuating the fine-grained features of critical force-generating phases, leading to inadequate representational capacity for the rhythmic and force-generation mechanisms of martial arts movements, and consequently limiting the discriminative accuracy at action boundaries [19, 20]. At the engineering deployment level, high-precision graph convolutional models rely on complex network architectures and massive parameter counts, incurring substantial computational overhead and high inference latency, which renders them incompatible with real-time deployment scenarios on mobile and edge devices [21, 22]. Although lightweight networks can effectively reduce computational complexity, a general degradation in feature representational capacity is commonly observed, making it difficult to balance recognition accuracy and inference efficiency in high-precision specialized martial arts recognition tasks. The inherent trade-off between accuracy and real-time performance has thus far remained unresolved [23].

To address the aforementioned technical bottlenecks in the field of martial arts action recognition, an end-to-end real-time action recognition and analysis framework tailored to the motion characteristics of specialized martial arts is developed. A dynamic deformable adaptive graph convolution module is designed, integrating anatomical priors with data-driven mechanisms to achieve dynamic adaptive optimization of joint topology and receptive fields, thereby resolving the issue of insufficient adaptability in fixed spatial modeling. Grounded in the phased motion regularities of martial arts, a semantics-driven multi-scale temporal phase perception module is constructed to enable adaptive action phase segmentation and differentiated temporal feature extraction, strengthening the model's capacity for modeling the rhythmic and force-generation characteristics of martial arts movements. Model scale compression is achieved through lightweight network optimization and knowledge distillation strategies, coupled with a multimodal cross-attention fusion mechanism to enhance recognition robustness in complex scenarios. A specialized martial arts action dataset is concurrently constructed, and the effectiveness and superiority of the proposed methodology are thoroughly validated through multi-dimensional comparative experiments and ablation studies.

The overall chapter structure of this study is arranged as follows. In Section 2, the overall architecture of the proposed system and the design principles of each core module are elaborated in detail. In Section 3, the experimental setup and dataset information are presented, and comprehensive model performance validation is conducted through comparative experiments, ablation studies, real-time performance analysis, and robustness analysis. In Section 4, the research findings of the entire study are summarized, the limitations of the current research are analyzed, and future research directions are discussed.

## 2. METHODOLOGY

### 2.1 Overall system framework

An end-to-end real-time martial arts action recognition and analysis system is developed, targeting the inherent characteristics of martial arts movements—namely, dynamically variable joint coordination, pronounced hierarchical temporal rhythms, and large instantaneous motion amplitudes—to overcome the homogenized processing limitations of fixed spatiotemporal modeling in conventional graph convolutional models. A hierarchical inference architecture is designed, incorporating spatially dynamic adaptation, temporally semantic stratification, and modality-adaptive enhancement. Continuous Red-Green-Blue (RGB) video is employed as the raw input, from which human joint coordinate information is extracted via a lightweight pose estimation network. Through temporal noise reduction and multi-dimensional motion feature reconstruction strategies, standardized skeletal spatiotemporal graph data are constructed, encompassing spatial positions, motion velocities, and skeletal geometric attributes, thereby providing a fundamental input for the fine-grained characterization of complex martial arts motion patterns. The entire architecture operates according to the progressive logic of spatial topology optimization, temporal feature mining, multimodal robustness enhancement, and lightweight inference output. Each functional module is deeply adapted to the specific motion mechanisms of martial arts, while simultaneously accommodating both feature representation precision and the

real-time requirements of edge-device deployment.

To quantitatively define the complete inference pipeline of the system, a continuous input video sequence of length $T$ is considered, and a multi-level feature mapping model from raw visual data to action recognition results is established. The overall computational process is formulated as:

$$\hat{y}=F_{cls}\left(F_{fuse}\left(F_{MSTM}\left(F_{DAG}\left(F_{pose}(I_{1:T})\right)\right)\right)\right) \tag{1}$$

where, $I_{1:T}$ denotes the raw input video sequence comprising $T$ frames; $F_{pose}$ is designated as the pose preprocessing function, responsible for skeletal keypoint detection, temporal filtering, and high-dimensional feature construction; $F_{DAG}$ is defined as the deformable adaptive graph convolution operation, which serves to dynamically rectify joint topological structures and neighborhood receptive fields, thereby fitting the transiently varying joint coordination relationships inherent in martial arts movements; $F_{MSTM}$ is specified as the multi-scale temporal modeling operation, which performs action temporal phase segmentation and differentiated scale feature extraction to mine the phased rhythmic semantics of martial arts actions; $F_{fuse}$ is designated as the cross-modal fusion operation, which adaptively fuses RGB appearance features with skeletal structural features according to scene complexity, thereby enhancing the model's anti-interference capability; and $F_{cls}$ is defined as the lightweight classification operation, which maps high-dimensional spatiotemporal features to the action category space and outputs the final recognition result. This modular architecture enables a decoupled design between feature modeling and engineering optimization, allowing flexible switching between unimodal and multimodal inference modes, and effectively balancing recognition accuracy and inference latency.

### 2.2 Skeletal sequence extraction via lightweight pose estimation

Video frame-level human joint perception is accomplished through a lightweight pose estimation network. To address the issues of keypoint jitter and temporal offset caused by the high motion speed and large limb deformation amplitudes characteristic of martial arts movements, skeletal data preprocessing and denoising optimization are performed. For an input video sequence of length $T$, two-dimensional coordinate information for 17 core human joints is extracted frame by frame, and a frame-level static skeletal topology is constructed. The skeletal graph corresponding to the $t$-th frame is defined as $G_t=(V,E)$, where $V$ denotes the set of human joint nodes, and $E$ denotes the fixed topological edge relationships established according to anatomical connection rules, serving to constrain the basic human limb structural associations. The raw detected joint coordinates exhibit significant temporal noise that disrupts the spatiotemporal continuity of continuous martial arts motion. A Kalman filtering algorithm is therefore employed to perform temporal smoothing correction of joint trajectories, suppressing high-frequency jitter through an iterative mechanism of state prediction and observation correction. The specific update equations are formulated as:

$$\hat{x}_{t|t}=\hat{x}_{t|t-1}+K_t(z_t-H\hat{x}_{t|t-1}) \tag{2}$$

$$P_{t|t}=(I-K_tH)P_{t|t-1} \tag{3}$$

where, $\hat{x}_{t|t-1}$ denotes the prior predicted state of the joint coordinates, $\hat{x}_{t|t}$ denotes the filtered posterior state, $z_t$ denotes the raw observed coordinates from pose estimation, $H$ denotes the fixed observation matrix, $K_t$ denotes the adaptive Kalman gain, and $P$ denotes the state covariance matrix, which is used to dynamically correct state estimation errors, thereby effectively ensuring the temporal stability and trajectory smoothness of skeletal keypoints under high-speed martial arts motion.

To precisely characterize the complex limb coordination and deformation properties of martial arts movements, high-dimensional node representations are constructed from the smoothed joint coordinate sequences through the integration of kinetic features and skeletal geometric features, thereby overcoming the limited motion representation capacity of raw coordinate features alone. Joint motion velocities and accelerations are respectively computed from consecutive frame joint position differences to quantify the instantaneous force generation and dynamic limb variation characteristics of martial arts movements. The computations are formulated as:

$$\Delta v_i^t=v_i^t-v_i^{t-1} \tag{4}$$

$$\Delta^2 v_i^t=v_i^{t+1}-2v_i^t+v_i^{t-1} \tag{5}$$

where, $\Delta v_i^t$ and $\Delta^2 v_i^t$ denote the motion velocity and acceleration, respectively, of the $i$-th joint at the $t$-th frame. Concurrently, skeletal geometric representations are constructed by defining the bone vector between adjacent joints as $b_{ij}^t=v_i^t-v_j^t$, from which bone Euclidean lengths and normalized directional features are derived through vector computation, thereby providing a complete description of limb posture deformation patterns. The raw joint coordinates, detection confidences, motion velocities, motion accelerations, bone lengths, and bone directions are integrated to form 11-dimensional node feature vectors. A standardized skeletal spatiotemporal tensor of dimension $\mathrm{R}^{T\times N\times C}$ is ultimately constructed, providing high-precision and high-robustness foundational data for subsequent dynamic graph convolution modeling and temporal feature mining.

### 2.3 Deformable adaptive graph convolution module

Conventional spatial graph convolution relies on fixed human anatomical topology to accomplish node information aggregation, with feature iteration across network layers depending on constant adjacency relationships for message passing, rendering it inadequate for adapting to the unique instantaneous joint coordination variation characteristics of martial arts movements. The feature update formulation of standard graph convolution can be uniformly expressed as:

$$H^{(l+1)}=\sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{6}$$

where, $H^{(l)}$ denotes the input node features at the $l$-th network layer, $W^{(l)}$ denotes the learnable weight matrix between layers, $\tilde{A}$ denotes the fixed anatomical adjacency matrix incorporating self-loop structures, $\tilde{D}$ denotes the corresponding degree matrix, and $\sigma$ denotes the nonlinear activation function. The fixed topological structure is capable only of characterizing static physiological connectivity relationships of the human body, failing to accommodate the irregular joint coordination

patterns under dynamic martial arts movements such as leaping, twisting, and explosive force exertion. To address this problem, a deformable adaptive graph convolution module is designed, integrating anatomical prior structures, data-driven association learning, and dynamic receptive field sampling mechanisms to enable adaptive evolutionary modeling of skeletal spatial topology, thereby accommodating the complex and variable limb motion patterns of martial arts. Figure 1 illustrates the network architecture of the proposed deformable adaptive graph convolution module.

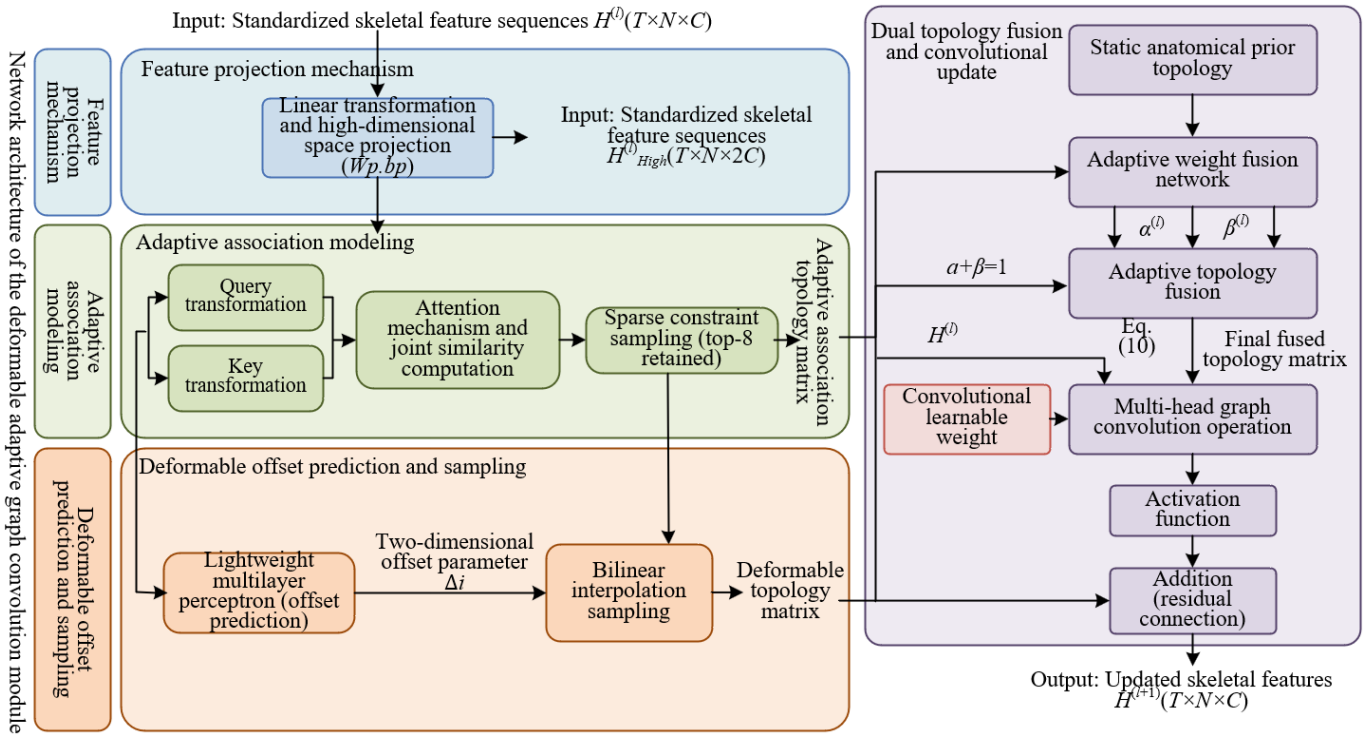


**Figure 1.** Network architecture of the deformable adaptive graph convolution module

To enhance the representational discriminability of joint motion features and provide high-precision feature support for subsequent dynamic topology learning, high-dimensional space projection transformation is performed on the raw node features. Low-dimensional base features are migrated to a high-dimensional latent space through linear mapping, fully exploiting the motion discrepancy information across different joints. The feature projection process is defined as:

$$\tilde{H}^{(l)}=H^{(l)}W_p+b_p \tag{7}$$

where, $W_p$ denotes the dimensional transformation weight matrix and $b_p$ denotes the bias parameter. The latent space dimension is uniformly set to twice the dimension of the original features, thereby expanding the feature representation space without introducing redundant computation, effectively enhancing the model's capacity to capture subtle limb deformations and coordination discrepancies in martial arts movements, and providing a robust feature foundation for subsequent joint similarity computation and offset prediction.

Based on the high-dimensional projected features, an attention mechanism is introduced to accomplish adaptive association modeling among joint nodes, freeing the model from the constraints of fixed anatomical connections and enabling action-specific joint coordination relationships to be learned autonomously from motion data. The similarity weight between joints is computed as:

$$A^{(i,j)}_{adapt}=softmax_j\left(\frac{(\tilde{H}^{(l)}_i W_q)(\tilde{H}^{(l)}_j W_k)^{\top}}{\sqrt{d_h}}\right) \tag{8}$$

where, $W_q$ and $W_k$ denote the projection matrices for the query and key vectors, respectively, and $d_h$ denotes the dimension of the high-dimensional latent space. To balance modeling accuracy and computational complexity, a sparse constraint is imposed on the adaptive similarity matrix, retaining the eight most similar neighboring nodes for each node while uniformly zeroing out all other association weights. This sparse sampling strategy effectively filters out interference from irrelevant joint associations, enabling precise focus on the genuinely coupled limb nodes during martial arts motion execution, and thereby constructs a dynamic association topology that conforms to the specific motion patterns of the sport.

To accommodate the large-amplitude and instantaneous limb motion characteristics of martial arts, a deformable offset prediction mechanism is introduced into the module, enabling dynamic adaptive adjustment of the graph convolution receptive field. A lightweight multilayer perceptron is employed to predict two-dimensional offset parameters based on the real-time motion features of each joint, thereby accomplishing dynamic correction of the neighborhood sampling positions. The offset computation process is formulated as:

$$\Delta^{(l)}_i=MLP_{offset}(\tilde{H}^{(l)}_i) \tag{9}$$

Based on the predicted offsets, bilinear interpolation sampling is performed on the adaptive topology matrix to obtain the deformable topology matrix. To integrate static structural priors with dynamic motion features, adaptive fusion of the dual topologies is achieved through learnable weights:

$$A_{final}^{(l)}=\alpha^{(l)}A_{anat}+\beta^{(l)}A_{deform}^{(l)},\alpha^{(l)}+\beta^{(l)}=1 \quad (10)$$

The fusion weights are adaptively generated by a learnable network based on global features, enabling the shallower layers of the network to prioritize anatomical topology for ensuring structural rationality, while the deeper layers emphasize dynamic topology for mining sport-specific motion features, thereby achieving differentiated topological modeling across different network layers.

Leveraging the fused dynamic topology matrix, the module accomplishes the final graph convolution feature aggregation and iterative update. Cross-node information interaction is realized after normalization of the topology matrix, with the complete convolution operation defined as:

$$H^{(l+1)}=\sigma\left(\tilde{D}_{final}^{-1/2}\tilde{A}_{final}^{(l)}\tilde{D}_{final}^{-1/2}H^{(l)}W^{(l)}\right) \quad (11)$$

where, $A^{\sim(l)}_{final}$ denotes the fused adjacency matrix with self-loops incorporated, and $\tilde{D}_{final}$ denotes the corresponding normalized degree matrix. To preserve shallow-layer effective features and mitigate the gradient attenuation problem in deeper networks, a residual connection is introduced to fuse the convolution output features with the original input features. This modeling approach enables dynamic adjustment of the information propagation range and weight allocation among joints according to the force-generation rhythm, motion amplitude, and limb coordination patterns of martial arts movements, completely breaking through the representational limitations of fixed topology in traditional graph convolution and achieving fine-grained modeling of complex spatial motion features in martial arts.

### 2.4 Martial-arts-semantics-driven multi-scale temporal modeling and phase perception module

Spatial graph convolution accomplishes dynamic association modeling in the spatial dimension of skeletal data, effectively characterizing the instantaneous coordination patterns of martial arts limbs, while temporal modeling is responsible for characterizing the rhythmic regularities and semantic hierarchies of complete actions as they evolve over time. Conventional temporal modeling methods generally employ fixed-receptive-field convolutional kernels to uniformly process temporal sequences, rendering them incapable of adapting to the inherent phased motion characteristics of martial arts actions. Martial arts routine actions exhibit clear state divisions, with significant differences in limb motion speed, deformation range, and force-generation mechanisms across distinct motion phases. The uniform temporal feature extraction approach fails to differentiate the semantic contributions of each phase, resulting in insufficient representational capacity of the model for action boundaries and critical force-generation features. A martial-arts-semantics-driven multi-scale temporal modeling and phase perception module is constructed, achieving hierarchical fine-grained modeling of temporal features through data-driven automatic phase division, scale-adaptive temporal feature extraction, and adaptive phase weight allocation, thereby precisely matching the temporal evolution patterns specific to martial arts movements. Figure 2 illustrates the flowchart of the proposed multi-scale temporal modeling and phase perception module.

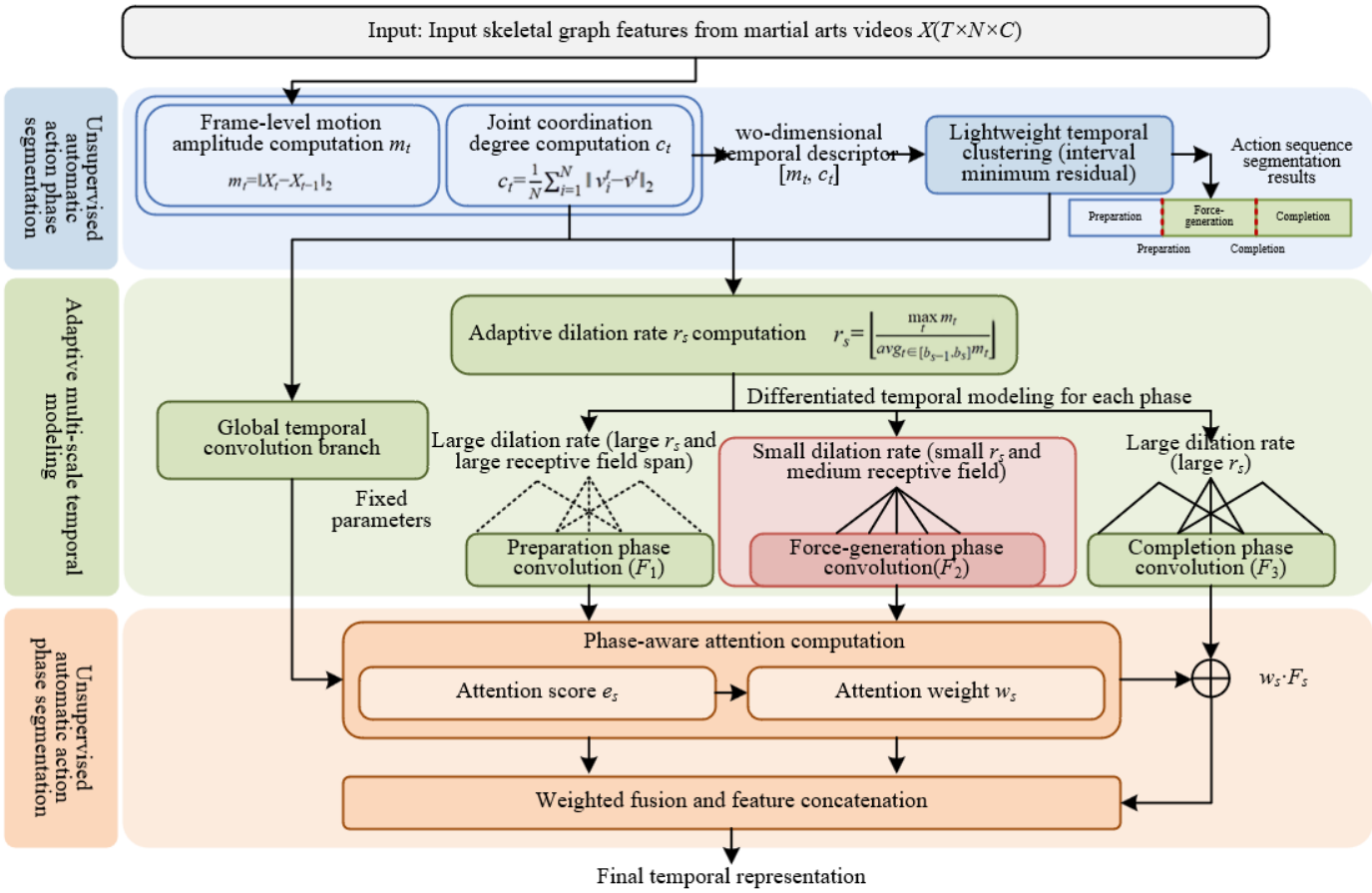


**Figure 2.** Flowchart of the multi-scale temporal modeling and phase perception module

The module first constructs frame-level motion representations to accomplish unsupervised automatic action phase segmentation, thereby overcoming the limitations of manually defined temporal interval division. Based on the skeletal temporal feature matrix of dimension $T \times N \times C$, the single-frame motion amplitude feature and joint coordination feature are respectively computed to quantify the temporal motion states. The computation formulas are as follows:

$$m_t = \| X_t - X_{t-1} \|_2 \tag{12}$$

$$c_t = \frac{1}{N} \sum_{i=1}^{N} \| v_i^t - \bar{v}^t \|_2, \bar{v}^t = \frac{1}{N} \sum_{i=1}^{N} v_i^t \tag{13}$$

where, $m_t$ denotes the overall motion variation amplitude at the $t$-th frame, used to characterize instantaneous motion intensity; $X_t$ and $X_{t-1}$ denote the skeletal features of two consecutive frames, respectively; $c_t$ denotes the intra-frame joint motion coordination degree, used to characterize limb coordination consistency; $N$ denotes the total number of human joints; $v_i^t$ denotes the motion vector of the $i$-th joint at the $t$-th frame; and $\bar{v}^t$ denotes the average motion vector of all joints within the frame. The two types of features are fused to construct a two-dimensional temporal descriptor, which is fed into a lightweight temporal clustering model. With the minimization of interval feature residual as the optimization objective, the complete action sequence is partitioned into three stable semantic phases, corresponding respectively to the preparation, force-generation, and completion processes of martial arts actions. The ordered temporal segmentation boundaries are output, providing a foundational basis for differentiated temporal modeling.

Based on the temporal phase segmentation, multi-scale differentiated feature extraction is performed on distinct motion phases using adaptive dilation rate temporal convolution, accommodating the motion rhythm variations across phases. The convolution dilation rate is dynamically determined according to the motion intensity within each interval, enabling adaptive adjustment of the receptive field. The dilation rate computation is formulated as:

$$r_s = \left\lfloor \frac{\max_t m_t}{avg_{t \in [b_{s-1}, b_s]} m_t} \right\rfloor \tag{14}$$

where, $r_s$ denotes the convolution dilation rate for the $s$-th action phase; $\max_t m_t$ denotes the maximum motion amplitude across the entire sequence; $avg_{t \in [b_{s-1}, b_s]} m_t$ denotes the average motion amplitude within the current phase; and $b_{s-1}$ and $b_s$ denote the start and end segmentation boundaries of the current temporal phase. Based on the adaptive dilation rate, the temporal feature extraction process for each phase is defined as:

$$F_s(\tau) = \sum_{p=0}^{k_s - 1} W_s(p) \cdot X_{b_{s-1} + \tau \cdot r_s \cdot p} \tag{15}$$

where, $k_s$ denotes the phase-specific convolution kernel size, $W_s(p)$ denotes the learnable convolution weight parameters, and $\tau$ denotes the temporal position index. A small dilation rate convolution is configured for the high-dynamic force-generation phase to capture fine-grained instantaneous motion features, while a large dilation rate convolution is configured for the static transition and completion phases to mine long-range global motion trends. Simultaneously, a global temporal convolution branch with fixed parameters is established to extract unified contextual features from the complete sequence, achieving complementary fusion of local phase-specific features and global temporal features.

To further enhance the dominant role of features from critical motion phases, a phase-aware attention mechanism is introduced into the module to adaptively quantify the semantic contributions of distinct temporal phases. The computation process for attention scores and normalized weights is formulated as:

$$e_s = MLP_{att}(F_s) \tag{16}$$

$$w_s = \frac{exp(e_s)}{\sum_{k=1}^{K} exp(e_k)} \tag{17}$$

where, $e_s$ denotes the raw attention score for the $s$-th phase feature, obtained through mapping by a lightweight fully connected network; $w_s$ denotes the normalized phase-wise attention weight; and $K$ denotes the total number of action phases. The final temporal features are obtained through weighted fusion and feature concatenation as:

$$F_{temp} = Concat\left( \sum_{s=1}^{K} w_s \cdot F_s, F_{global} \right) \tag{18}$$

where, $F_{temp}$ denotes the final temporal representation output by the module, and $F_{global}$ denotes the global temporal convolution features. This architecture automatically elevates the feature weights of critical phases according to the force-generation characteristics of martial arts movements, while attenuating feature interference from ineffective transition phases, thereby effectively compensating for the deficiencies of fixed temporal modeling methods. Subsequent ablation experiments further validate the incremental gains achieved by phase segmentation, multi-scale convolution, and the attention mechanism.

### 2.5 Lightweight deployment strategy

To enable real-time deployment of the martial arts action recognition model on edge terminals, a hierarchical lightweight optimization system is constructed from three dimensions: spatial structure optimization, temporal data reduction, and model knowledge transfer. The aforementioned dynamic graph convolution and temporal perception modules enable high-precision specialized action modeling; however, the complex spatiotemporal interactive computations incur substantial parameter counts and computational overhead, rendering them difficult to satisfy the low-latency inference requirements of mobile devices. Through the collaborative application of multi-level lightweight strategies, the computational load of the model is compressed, significantly reducing inference cost while fully preserving the semantic information of martial arts force-generation characteristics, temporal rhythms, and joint coordination, thereby achieving an optimal trade-off between recognition accuracy and computational efficiency.

A grouped graph convolution mechanism is introduced to optimize the computational complexity of spatial modeling, partitioning all joints into independent subgroups according to human motor functional characteristics, thereby

circumventing redundant computations arising from global node-wise interactions. The 17 human joints are divided into three functional units—upper limbs, lower limbs, and torso—with each joint subgroup independently performing graph convolution feature extraction, and the output features from all groups are finally concatenated to accomplish spatial representation fusion. The computation process is defined as:

$$H_{group}^{(l+1)}=Concat\begin{pmatrix} GConv_1(H_1^{(l)}),GConv_2(H_2^{(l)}), \\ \ldots,GConv_G(H_G^{(l)}) \end{pmatrix} \tag{19}$$

where, $G$ denotes the total number of joint groups, $GConv_g$ denotes the independent graph convolution operation corresponding to the $g$-th group, and $H^{(l)}{}_g$ denotes the node features of the $g$-th group of joints at the $l$-th network layer. To circumvent the issue of inter-group feature isolation caused by independent group convolutions, a group shuffling mechanism is introduced to accomplish cross-group information exchange:

$$H_{shuffle}^{(l+1)}=Shuffle(H_{group}^{(l+1)}) \tag{20}$$

This architecture reduces the computational complexity of global graph convolution from $O(N^2d)$ to $O(\Sigma^G{}_{g=1}N^2{}_gd)$, where $N_g$ denotes the number of joints in a single group and $d$ denotes the feature dimension, thereby compressing the overall spatial computational cost to approximately one-third of the original while ensuring the learnability of cross-limb coordination features.

To address the non-uniform temporal motion characteristics of martial arts movements, a motion-amplitude-adaptive temporal sparse sampling strategy is designed, replacing the fixed-interval sampling approach and effectively eliminating redundant frames from quiescent phases. Based on the frame-level motion amplitude features, the global motion mean and median values across the entire sequence are statistically computed, and the temporal sampling stride is adaptively determined. The specific computation is formulated as:

$$\delta = \max\left(1, \left\lfloor \frac{M}{median(m_t)} \right\rfloor\right) \tag{21}$$

where, $\delta$ denotes the adaptive sampling stride, $M$ denotes the average motion amplitude of the complete video sequence, and $median(m_t)$ denotes the median value of motion amplitudes across all frames. Dense sampling with unit stride is employed for force-generation phases with motion amplitudes above the median value, preserving fine-grained temporal features of instantaneous explosive actions, while sparse sampling with large strides is applied to low-motion-amplitude transition and completion phases. The effective sequence length after sampling is statistically obtained through a binary indicator function. This strategy compresses the temporal data volume to below 70% of the original without losing critical action semantics, substantially reducing the computational overhead of temporal modeling.

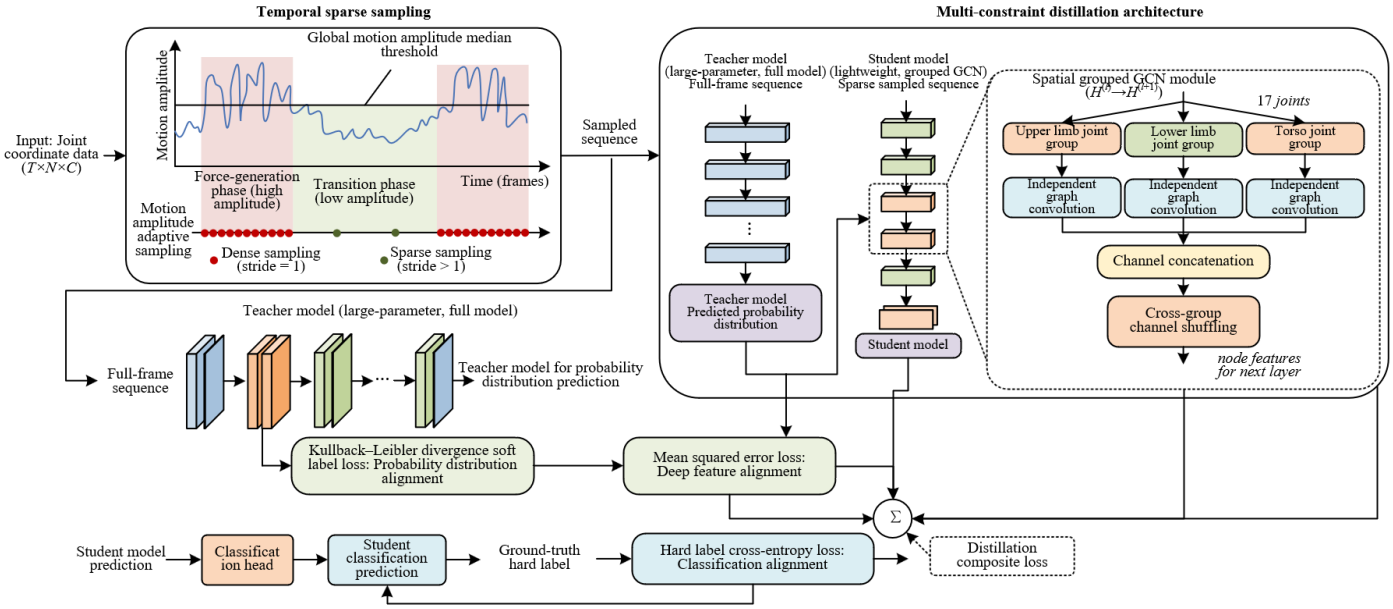


**Figure 3.** Model lightweighting and knowledge distillation framework for edge deployment

To compensate for the feature representational degradation caused by structural compression and sparse sampling, a multi-constraint knowledge distillation strategy is introduced to transfer the specialized motion modeling knowledge of the high-precision large model to the lightweight student network. Figure 3 illustrates the model lightweighting and knowledge distillation framework for edge deployment. The complete network with 15 M parameters is adopted as the teacher model, and the lightweight compressed network is adopted as the student model. A composite distillation loss function is constructed, integrating hard-label classification loss, feature matching loss, and soft-label probability loss:

$$L_{KD}=L_{CE}(y_S,y_{true})+\lambda_1 L_{MSE}(f_S(X),f_T(X))+\lambda_2 L_{KL}(p_S^\tau,p_T^\tau) \tag{22}$$

where, $L_{CE}$ denotes the cross-entropy loss, which enforces hard supervision learning from ground-truth labels; $L_{MSE}$ denotes the mean squared error loss, which constrains the alignment of deep feature distributions between the student and teacher networks; $L_{KL}$ denotes the Kullback–Leibler divergence loss, which achieves optimization of the predicted probability distribution fitting; $f_S$ and $f_T$ denote the feature

mapping functions of the student and teacher networks, respectively; and $p_S^\tau$ and $p_T^\tau$ denote the predicted probabilities normalized with the temperature coefficient, where the temperature parameter $\tau$ is used to smooth the probability distribution and mine fine-grained category differences. The weight coefficients $\lambda_1$ and $\lambda_2$ respectively regulate the loss contributions of feature distillation and probability distillation. Through multi-dimensional knowledge constraints, the lightweight model is enabled to acquire the teacher network's fine-grained modeling capabilities for martial arts joint coordination and temporal force generation, effectively addressing the accuracy degradation problem inherent in lightweight models.

### 2.6 Red-green-blue-skeletal cross-attention fusion module

Single-modality skeletal representations and RGB visual representations each possess inherent representational deficiencies in martial arts action recognition tasks. Skeletal temporal features stably characterize limb topological coordination patterns and are insensitive to illumination and viewpoint variations, yet fail to capture subtle limb morphology and texture differences in martial arts movements. RGB visual features preserve rich appearance details, but are highly susceptible to scene background interference, limb occlusion, and other factors, making precise focus on the moving subject difficult. To achieve deep complementarity between the two modalities, a bidirectional RGB-skeletal cross-attention fusion module is constructed, establishing a dual-branch adaptive information interaction mechanism to accomplish synergistic enhancement of visual appearance features and human structural features. Lightweight networks and the aforementioned spatiotemporal modeling networks are respectively employed for dual-branch feature extraction. Temporal visual features of dimension $T \times d_{RGB}$, denoted as $F_{RGB}$, are extracted frame by frame from video appearance information using MobileNetV3. Skeletal motion features of dimension $T \times d_{skel}$, denoted as $F_{skeleton}$, are output through the deformable adaptive graph convolution and temporal modeling network. Figure 4 illustrates the network architecture of the RGB-skeletal bidirectional cross-attention multimodal fusion module. Cross-modal guided enhancement is achieved through bidirectional cross-attention, with the specific interaction computations formulated as:

$$F_{RGB \leftarrow skeleton} = Softmax\left(\frac{F_{RGB} W_Q (F_{skeleton} W_K)^{\top}}{\sqrt{d}}\right) F_{skeleton} W_V \tag{23}$$

$$F_{skeleton \leftarrow RGB} = Softmax\left(\frac{F_{skeleton} W_Q' (F_{RGB} W_K')^{\top}}{\sqrt{d}}\right) F_{RGB} W_V' \tag{24}$$

where, $W_Q$, $W_K$, and $W_V$ denote the query, key, and value projection matrices for the vision-modality-guided branch, respectively; $W_Q'$, $W_K'$, and $W_V'$ denote the corresponding projection matrices for the skeleton-modality-guided branch; and $d$ denotes the unified feature mapping dimension. This interaction mechanism constrains the effective perceptual region of visual features using skeletal topological structures, suppressing noise interference from complex backgrounds, while simultaneously compensating for the detail deficiency in skeletal keypoint representations using RGB fine-grained appearance information, achieving bidirectional gain for both modalities.

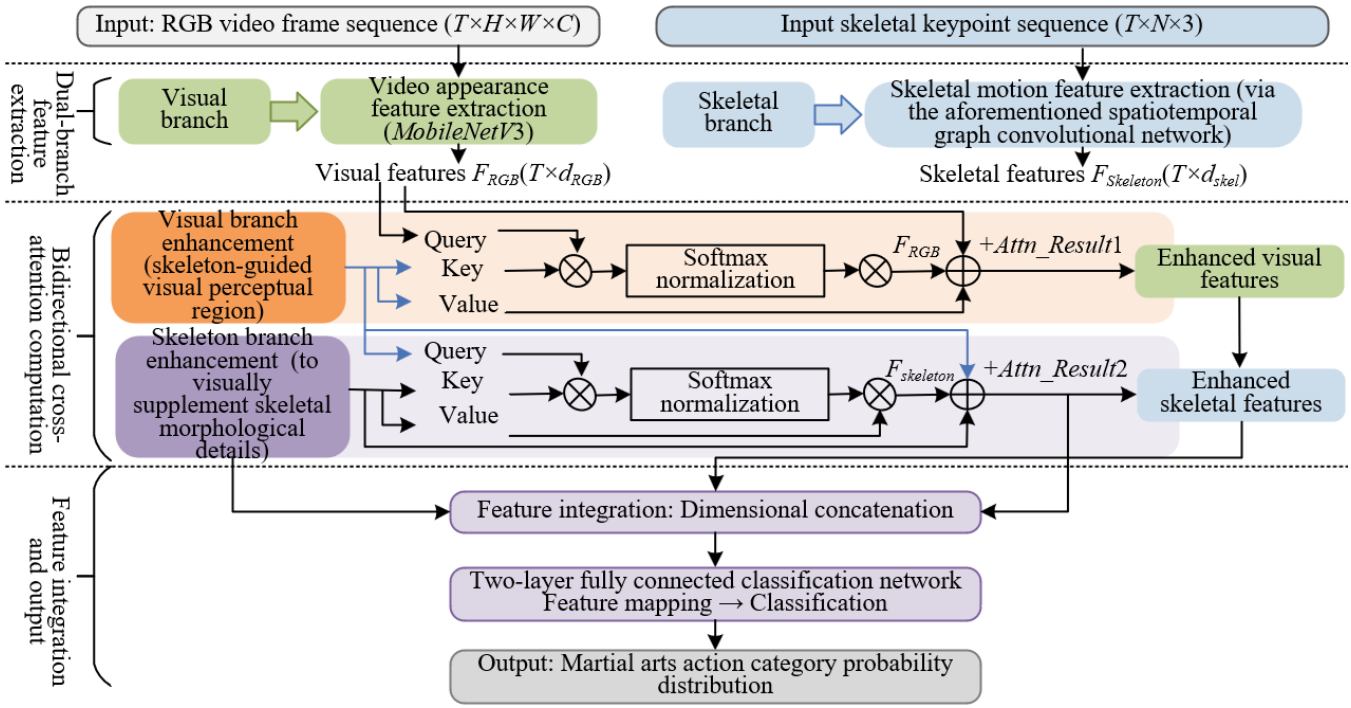


**Figure 4.** Network architecture of the Red-Green-Blue (RGB)-skeletal bidirectional cross-attention multimodal fusion module

Upon completion of the cross-modal attention interaction, the enhanced features from both branches are integrated through residual fusion and dimensional concatenation strategies to construct a highly robust comprehensive action representation. The fusion computation is formulated as:

$$F_{fuse} = Concat(F_{RGB} + F_{RGB \leftarrow skeleton},\ F_{skeleton} + F_{skeleton \leftarrow RGB}) \tag{25}$$

where, $F_{fuse}$ denotes the final multimodal fused feature. Through residual addition, the original effective features of each branch are preserved while cross-modal complementary information is incorporated, thereby maximizing the retention of both structural features and appearance details of martial arts movements. The fused features are fed into a two-layer fully connected classification structure to accomplish action category discrimination, with the normalized probability distribution output as:

$$\hat{y}=Softmax(F_{fuse}W_{cls}+b_{cls}) \tag{26}$$

where, $W_{cls}$ and $b_{cls}$ denote the learnable weights and bias terms of the classification layer, respectively, and $\hat{y}$ denotes the action recognition prediction output by the model. To accommodate the real-time inference requirements of edge devices, an adaptively switchable operation mechanism is configured for this module. The multimodal fusion strategy is enabled under adverse scenarios such as occlusion and complex backgrounds to ensure recognition accuracy, while the cross-modal interaction branch is disabled under simple scenarios to reduce computational load, thereby achieving a dynamic balance between model robustness and real-time performance.

## 3. EXPERIMENTAL DESIGN AND RESULTS

To systematically validate the effectiveness, superiority, and engineering practicality of the proposed martial arts action recognition and analysis system, multi-dimensional comparative experiments, module ablation studies, real-time performance tests, and robustness validation experiments were conducted. The experiments encompassed both general-purpose human action benchmark datasets and a self-constructed specialized martial arts dataset, with quantitative evaluations performed across multiple dimensions, including recognition accuracy, model complexity, inference real-time performance, and scene anti-interference capability. The design rationality of each core module and the comprehensive performance advantages of the overall framework were comprehensively demonstrated through these evaluations.

### 3.1 Experimental setup

Model training and performance evaluation were conducted using four public benchmark datasets and a self-constructed specialized martial arts dataset, encompassing validation of both general-purpose action recognition capability and martial arts-specific adaptability. The NTU RGB+D 60 dataset comprised 60 categories of daily human actions, with a total of 56,000 video samples, constructed through a multi-view acquisition scheme, and two mainstream evaluation protocols—cross-subject and cross-view—were adopted. The NTU RGB+D 120 dataset served as its extended version, covering 120 action categories, with performance evaluation conducted under cross-subject and cross-setup protocols. The Kinetics-Skeleton dataset contained 400 categories of complex human actions, with a cumulative total of 300,000 video clips, from which skeletal keypoint data were automatically extracted by pose estimation algorithms, making it suitable for large-scale general-purpose action recognition model performance validation. To accommodate the requirements of specialized martial arts research, the Wushu-Action self-constructed martial arts dataset was built, encompassing 10 classic martial arts routine action categories, including Tai Chi, Changquan, and Nanquan. Multi-session recordings were performed by 20 professional martial arts athletes, with each sample collected five times, ultimately resulting in a specialized dataset comprising 10,000 valid video samples. All samples were acquired in standard indoor training scenarios, with a unified resolution of 1920 × 1080 and a frame rate of 30 fps, ensuring the standardization and uniformity of data collection.

A comprehensive evaluation system was constructed using multi-dimensional quantitative metrics. At the recognition accuracy level, Top-1 classification accuracy was selected as the core evaluation metric to measure the model's action classification discrimination capability. At the model efficiency level, parameter count and floating-point operations were employed to quantify model complexity, while inference frames per second and per-frame inference latency were used to quantify real-time deployment performance. At the model robustness level, occlusion interference and viewpoint shift experiments were conducted to quantify the degree of performance degradation under complex scenarios, enabling comprehensive assessment of the model's environmental adaptability.

All experiments were implemented and trained using the PyTorch deep learning framework. The pose estimation module adopted a lightweight OpenPose variant optimized with MobileNetV2, ensuring real-time keypoint extraction at the front end. Model training was performed using the Adam optimizer, with an initial learning rate of 0.001, coupled with a cosine annealing learning rate scheduling strategy for dynamic learning rate decay. The batch size was set to 32, and the total number of training epochs was 120, with the hardware environment consisting of dual NVIDIA RTX 3090 graphics processing units. The model distillation hyperparameters followed the settings established in the methodology section, with the temperature coefficient set to 4, and the loss weight coefficients configured as 0.1 and 0.5, respectively. The network input was uniformly fixed to a temporal length of 64 frames and 17 human skeletal keypoint nodes.

### 3.2 Comparative experiments

To validate the performance advantages of the proposed algorithm over existing mainstream skeleton-based action recognition methods, classic and state-of-the-art graph convolutional recognition models from recent years were selected for horizontal comparative experiments, encompassing both lightweight models and high-precision large models, with tests conducted on both general-purpose benchmark datasets and the specialized martial arts dataset.

As shown in Table 1, the proposed model achieves optimal overall performance on the NTU series of general-purpose datasets. Under the NTU-60 cross-subject and cross-view evaluation protocols, recognition accuracies of 93.9% and 96.5% are achieved, respectively, surpassing the recent state-of-the-art deformable adaptive graph convolutional network (DA-GCN) model by 0.4 and 0.3 percentage points, respectively. On the more challenging NTU-120 dataset, cross-subject and cross-setup accuracies of 91.2% and 92.1% are attained, with the performance advantage consistently maintained. Compared with various competing models, the proposed method, with a lightweight parameter count of only 4.8 M, overcomes the inherent problems of parameter

redundancy in high-precision models and insufficient accuracy in lightweight models. Through the dynamic topology modeling and temporal semantic perception mechanisms, general-purpose action recognition accuracy is improved while model complexity is substantially reduced, achieving an optimal trade-off between accuracy and efficiency.

**Table 1.** Performance comparison with mainstream methods on the NTU RGB+D dataset

| Method | Year | Parameters (M) | NTU-60 X-Sub (%) | NTU-60 X-View (%) | NTU-120 X-Sub (%) | NTU-120 X-Set (%) |
|---|---|---|---|---|---|---|
| ST-GCN | 2018 | 3.1 | 81.5 | 88.3 | 79.2 | 81.5 |
| 2s-AGCN | 2019 | 6.9 | 88.5 | 93.2 | 82.9 | 84.9 |
| MS-G3D | 2020 | 7.8 | 89.8 | 94.1 | 86 | 87.4 |
| CTR-GCN | 2021 | 5.6 | 90.6 | 94.6 | 87.3 | 88.7 |
| DeGCN | 2023 | 8.2 | 91.5 | 95.3 | 88.6 | 89.8 |
| DA-GCN | 2025 | 7.16 | 93.5 | 96.2 | 90.63 | 91.7 |
| LI-AGCN | 2025 | 0.18 | 90.03 | 94.1 | — | — |
| Proposed full model | — | 4.8 | 93.9 | 96.5 | 91.2 | 92.1 |

Note: ST-GCN = Spatial Temporal Graph Convolutional Network; 2s-AGCN = Two-Stream Adaptive Graph Convolutional Network; MS-G3D = Multi-Scale Graph 3D Convolutional Network; CTR-GCN = Channel-Wise Topology Refinement Graph Convolutional Network; DeGCN = Deformable Graph Convolutional Network; DA-GCN = Deformable Adaptive Graph Convolutional Network; LI-AGCN = Lightweight Initialization-Enhanced Adaptive Graph Convolutional Network

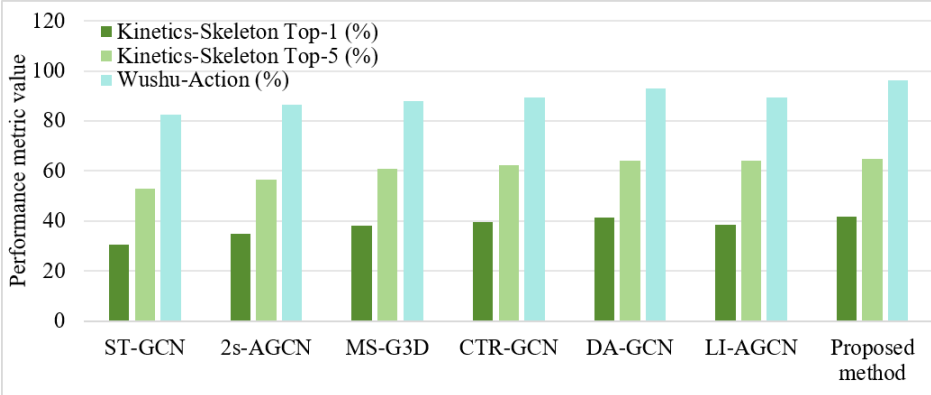


**Figure 5.** Performance comparison on Kinetics-Skeleton and the self-constructed martial arts dataset

Note: ST-GCN = Spatial Temporal Graph Convolutional Network; 2s-AGCN = Two-Stream Adaptive Graph Convolutional Network; MS-G3D = Multi-Scale Graph 3D Convolutional Network; CTR-GCN = Channel-Wise Topology Refinement Graph Convolutional Network; DA-GCN = Deformable Adaptive Graph Convolutional Network; LI-AGCN = Lightweight Initialization-Enhanced Adaptive Graph Convolutional Network

As shown in Figure 5, the results on both the large-scale general-purpose dataset and the specialized dataset demonstrate that the proposed model possesses excellent generalization capability and specialized adaptability. On the Kinetics-Skeleton dataset, Top-1 and Top-5 accuracies of 41.8% and 64.9% are achieved, respectively, representing steady improvements over the state-of-the-art DA-GCN model, thereby confirming that the multi-scale temporal modeling and dynamic graph convolution structure can be effectively adapted to large-scale, multi-category complex human actions. On the self-constructed Wushu-Action martial arts specialized dataset, a recognition accuracy of 96.2% is achieved by the proposed model, representing improvements of 3.4 percentage points over DA-GCN and 7.0 percentage points over the Lightweight Initialization-Enhanced Adaptive Graph Convolutional Network (LI-AGCN). The significant performance gaps indicate that existing general-purpose graph convolutional models are inadequate for adapting to the specialized characteristics of martial arts movements, namely dynamically variable joint coordination and prominently layered temporal rhythms. In contrast, the deformable adaptive graph convolution and phase-aware temporal module designed in this study are capable of precisely capturing the force-generation patterns and limb deformation characteristics of martial arts, thereby possessing irreplaceable modeling advantages in specialized motion recognition tasks.

### 3.3 Ablation studies

To clarify the independent contributions and synergistic effects of each core module, stepwise ablation experiments were conducted based on the baseline model, with separate validation of the overall module combination effect, the spatial modeling sub-module performance, and the temporal modeling sub-module performance. The experiments were simultaneously conducted on the self-constructed martial arts dataset and the NTU-60 dataset, ensuring the generality and specialized adaptability of the conclusions.

As clearly shown in Table 2, the incremental gains and synergistic effects of each module are demonstrated. With the classic spatial temporal graph convolutional network adopted as the baseline model, the recognition accuracy on the martial arts dataset is improved by 6.2 percentage points upon the sole introduction of the deformable adaptive graph convolution module, confirming that dynamic topology modeling effectively resolves the problem of fixed anatomical topology being inadequate for adapting to the dynamic joint coordination of martial arts movements. Upon the sole introduction of the multi-scale temporal phase perception module, the accuracy is improved by 3.9 percentage points,

validating the representational capacity of phased temporal modeling for the layered rhythmic characteristics of martial arts. When the two modules are used cooperatively, the overall accuracy improvement exceeds the superposition of individual module gains, exhibiting a significant super-additive synergistic characteristic, indicating that spatial dynamic modeling and temporal semantic modeling can form effective complementarity, jointly perfecting the spatiotemporal feature representation system for martial arts actions. Following the introduction of the lightweight knowledge distillation strategy, further improvement in model accuracy is observed, suggesting that knowledge distillation not only compresses model parameters but also provides regularization constraints through the prior knowledge of the teacher model, alleviating the feature degradation problem induced by lightweighting. The final full model, incorporating the multimodal cross-attention module, achieves optimal accuracy, fully validating the feature enhancement effect of RGB and skeletal dual-modality fusion under complex scenarios.

**Table 2.** Ablation experiments of each module on the self-constructed martial arts dataset

| Configuration | DAG-Conv | MSTM | Lightweighting | Multimodal | Wushu-Action (%) | NTU-60 X-Sub (%) |
|---|---|---|---|---|---|---|
| Baseline (ST-GCN) | × | × | × | × | 82.5 | 81.5 |
| +DAG-Conv | √ | × | × | × | 88.7 (+6.2) | 87.3 (+5.8) |
| +MSTM | × | √ | × | × | 86.4 (+3.9) | 85.1 (+3.6) |
| + Deformable Adaptive Graph Convolution + Multi-Scale Temporal Modeling | √ | √ | × | × | 92.1 (+9.6) | 90.8 (+9.3) |
| + Lightweighting (distillation) | √ | √ | √ | × | 93.5 (+11.0) | 92.2 (+10.7) |
| Full model | √ | √ | √ | √ | 96.2 (+13.7) | 93.9 (+12.4) |

Note: ST-GCN = Spatial Temporal Graph Convolutional Network; DAG-Conv = Deformable Adaptive Graph Convolution; MSTM = Multi-Scale Temporal Modeling

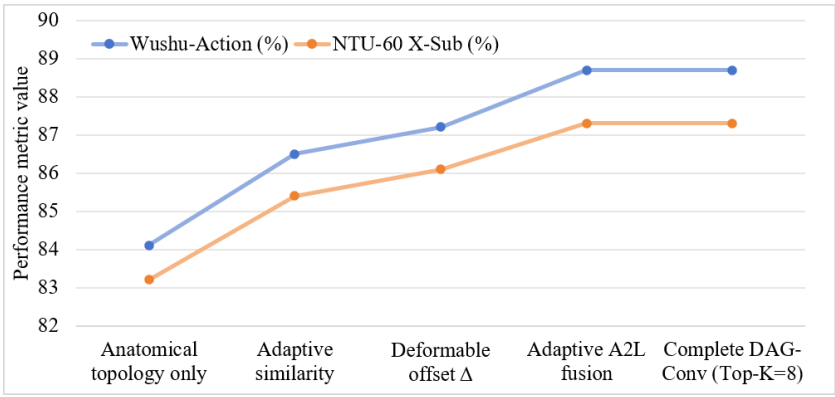


**Figure 6.** Internal ablation of the deformable adaptive graph convolution module

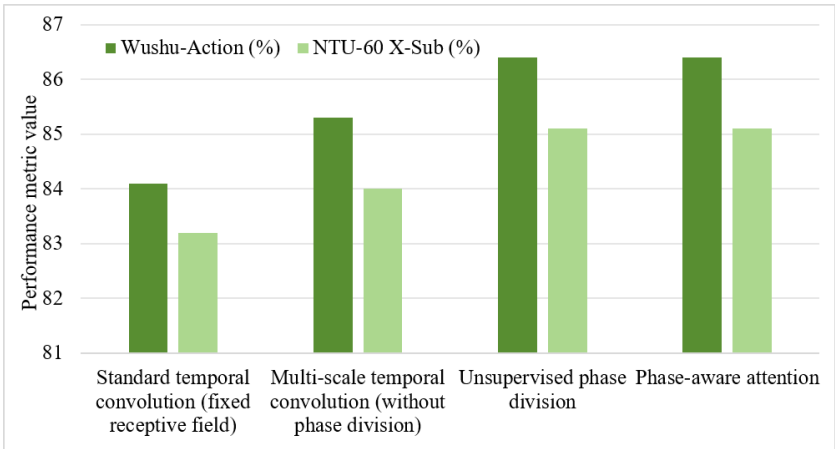


**Figure 7.** Internal ablation of the martial-arts-semantics-driven multi-scale temporal modeling module

Stepwise ablation of the internal sub-structures of the deformable adaptive graph convolution module is conducted in Figure 6, with the independent value of each component systematically decomposed. Limited performance is exhibited by the model relying solely on static anatomical topology, which is incapable of adapting to dynamic motion scenarios. Following the introduction of data-driven adaptive similarity learning, action-specific joint association features are autonomously learned by the model, with significant improvement observed in accuracy. With the introduction of deformable offsets, dynamic adjustment of the convolution receptive field is enabled, accommodating the large-amplitude instantaneous limb deformations of martial arts movements and further refining the precision of spatial feature

representation. Through the adaptive fusion module, dynamic weighting is employed to balance static structural priors and dynamic motion features, circumventing the limitations of fixed fusion ratios and maximizing the adaptability of spatial modeling. The cumulative gain achieved through the sequential addition of each sub-component fully validates the rationality and necessity of the deformable adaptive graph convolution module architectural design.

The performance gain patterns of each sub-component of the multi-scale temporal modeling module are presented in Figure 7. Limited feature extraction capability is exhibited by standard temporal convolution with fixed receptive fields, which is incapable of adapting to the differentiated temporal motion rhythms of martial arts. With the introduction of multi-scale convolution, temporal features at different granularities can be captured, yielding fundamental performance improvements. Through unsupervised action phase division, semantic segmentation of action temporal sequences is accomplished, with differentiated modeling strategies adopted for the preparation, force-generation, and completion phases, effectively addressing the deficiencies of homogenized temporal modeling and serving as the core factor driving temporal performance enhancement. With the phase-aware attention mechanism, adaptive reinforcement of feature weights for critical force-generation phases is enabled, further aligning with the force-generation mechanisms of martial arts movements and perfecting the temporal semantic representation system. The stepwise ablation results comprehensively validate the adaptability of the hierarchical temporal modeling strategy to specialized martial arts actions.

(a) Raw martial arts action input

(b) Skeletal temporal smoothing optimization

(c) Dynamic joint association enhancement


(d) Final action recognition and analysis overlay

**Figure 8.** Visualization of image processing and recognition implementation effects of the proposed method on martial arts action sequences

To validate the skeletal perception, dynamic association modeling, and action recognition capabilities of the proposed system under complex postures in real-world martial arts scenarios, an end-to-end implementation effect test is conducted using the broadsword technique side lunge stance action. As shown in Figure 8, complex postures are exhibited in the raw image, including large-amplitude lateral extension, low-center-of-gravity support, and crossed limbs with weapon-holding. Following temporal smoothing processing, stable alignment of key joints—including the head, shoulders, elbows, wrists, hips, knees, and ankles—with human anatomical positions is achieved, with continuous skeletal connections and no significant structural drift observed, indicating that front-end pose estimation and temporal correction effectively suppress keypoint jitter induced by high-speed motion. Through the dynamic joint association enhancement results, cross-joint information interaction among the weapon-holding arm, torso core, and bilateral lower-limb support chain is further established, providing a relatively complete characterization of the synergistic relationships among upper-limb swinging, torso rotation, and lower-limb stable support in the broadsword technique. In the final system output, the action is recognized as "Broadsword Technique – Side Lunge Stance," with a recognition confidence of 97.3% and an action completion of 94.6%, indicating that the proposed deformable adaptive graph convolution and phase-aware temporal modeling mechanisms are capable of simultaneously maintaining skeletal structural stability, joint association integrity, and category discrimination reliability under complex martial arts postures, thereby providing effective support for real-time action recognition, technical standardization assessment, and quantitative analysis of training processes for martial arts athletes.

### 3.4 Real-time performance analysis

To validate the deployability of the model on edge devices, real-time performance tests were conducted from three dimensions: model complexity, desktop inference speed, and edge device inference speed. The comprehensive deployment performance was compared against mainstream algorithms, while lightweight strategies were ablated step by step to verify the operational mechanisms of each optimization approach.

As shown in Table 3, dual advantages in inference speed and recognition accuracy are achieved by the proposed lightweight student network. With a parameter count of 4.8M and floating-point operations of only 7.6 G, inference rates of

156 fps are attained on the RTX 3090 desktop graphics processing unit and 42.5 fps on the Jetson Xavier NX edge device, with a per-frame inference latency of only 23.5 ms, fully satisfying the engineering requirements of real-time training analysis. Compared with the high-precision DA-GCN model, a 33% reduction in parameter count is achieved by the proposed method, with edge inference speed improved by nearly 4 times. Compared with the LI-AGCN model, the proposed method achieves essentially equivalent inference speed, while improving specialized martial arts recognition accuracy by 7.0 percentage points, effectively resolving the problem of insufficient feature representational capacity in lightweight models. Although higher recognition accuracy is possessed by the uncompressed teacher network, its substantial computational overhead and extremely low edge inference frame rate render it impractical for deployment, fully demonstrating the engineering necessity of the lightweight distillation strategy.

**Table 3.** Model complexity and inference speed comparison

| Method | Parameters (M) | Floating-Point Operations (G) | RTX 3090 (fps) | Jetson Xavier NX (fps) | Latency (ms)* |
|---|---|---|---|---|---|
| ST-GCN | 3.1 | 16.2 | 85.3 | 12.7 | 78.7 |
| 2s-AGCN | 6.9 | 37.5 | 42.1 | 6.8 | 147.1 |
| MS-G3D | 7.8 | 48.6 | 31.5 | 5.2 | 192.3 |
| DeGCN | 8.2 | 52.3 | 28.7 | 4.3 | 232.6 |
| DA-GCN | 7.16 | 41.8 | 38.2 | 6.1 | 163.9 |
| LI-AGCN | 0.18 | 1.2 | 156.3 | 35.2 | 28.4 |
| Proposed method (teacher network) | 15.2 | 68.5 | 18.6 | 3.1 | 322.6 |
| Proposed method (lightweight student network) | 4.8 | 7.6 | 156 | 42.5 | 23.5 |

Note: *Latency is reported as the average inference time (ms) per frame on the Jetson Xavier NX.
ST-GCN = Spatial Temporal Graph Convolutional Network; 2s-AGCN = Two-Stream Adaptive Graph Convolutional Network; MS-G3D = Multi-Scale Graph 3D Convolutional Network; DeGCN = Deformable Graph Convolutional Network; DA-GCN = Deformable Adaptive Graph Convolutional Network; LI-AGCN = Lightweight Initialization-Enhanced Adaptive Graph Convolutional Network

**Table 4.** Ablation comparison of different lightweight strategies (tested on Jetson Xavier NX)

| Configuration | Parameters (M) | Floating-Point Operations (G) | Frames per Second | Wushu-Action (%) |
|---|---|---|---|---|
| Full model (without lightweighting) | 15.2 | 68.5 | 3.1 | 97.1 |
| + Grouped graph convolution (G=3) | 8.6 | 32.4 | 8.7 | 95.8 |
| + Temporal sparse sampling | 8.6 | 18.5 | 18.3 | 95.2 |
| + Knowledge distillation | 4.8 | 7.6 | 42.5 | 96.2 |
| + Multimodal fusion (optional) | 5.6 | 9.2 | 36.8 | 96.8 |

The optimization effects of the lightweight strategies are validated step by step in Table 4. Through grouped graph convolution, spatial interaction computations are streamlined via limb function grouping, substantially compressing model size with only a minor sacrifice in accuracy. Through temporal sparse sampling, redundant frame data are adaptively pruned based on motion amplitude, effectively reducing the computational overhead of temporal modeling. Through knowledge distillation, not only is extreme model parameter compression accomplished, but the accuracy loss induced by lightweighting is also compensated through the transfer of specialized motion knowledge from the teacher model, achieving accuracy recovery and forming an effective accuracy compensation mechanism. The multimodal fusion module can be flexibly enabled or disabled according to application scenarios; when enabled, further improvement in recognition accuracy under complex scenarios is achieved with only a slight sacrifice in inference speed, thereby realizing a dynamic balance between model performance and deployment efficiency, and accommodating differentiated deployment requirements across various scenarios.

### 3.5 Robustness analysis

In real-world martial arts training scenarios, interference factors such as limb occlusion and viewpoint shifts are commonly prevalent. To validate the environmental adaptability of the model, joint occlusion robustness experiments and viewpoint shift robustness experiments were respectively conducted, with the stability of model performance under various types of interference quantitatively evaluated.

As shown in Figure 9, a declining trend in recognition accuracy with increasing joint occlusion rates is observed across all compared models, while significantly superior anti-interference capability is exhibited by the proposed model. Through the dynamic topology adaptive mechanism, the model relying solely on the skeletal branch is capable of completing inference based on remaining joint association features when partial joint information is missing, with the performance degradation amplitude being significantly lower than that of mainstream models. Through the RGB-skeletal cross-attention fusion, the full model is able to compensate for missing skeletal data using the visual modality, maintaining a recognition accuracy of 84.3% even under extreme 40% joint occlusion conditions, with an overall average performance decline of only 11.9%. This fully demonstrates the anti-occlusion advantage of the multimodal fusion mechanism, which can effectively address common scenario interferences in martial arts training, such as limb crossing and equipment occlusion.

The viewpoint perturbation experimental results presented in Figure 10 demonstrate that excellent viewpoint generalization capability is possessed by the proposed model. Under the extreme condition of ±60° large-angle viewpoint shifts, a recognition accuracy of 82.4% is still maintained by the skeletal branch model, with the performance degradation caused by viewpoint shifts being far lower than that of existing

mainstream algorithms. Through the incorporation of RGB visual texture features, the full multimodal model is able to compensate for the representational deviations induced by skeletal topology viewpoint shifts, with an overall average performance decline of only 8.7%. These results validate that the dynamic graph convolution and temporal modeling mechanisms are capable of learning viewpoint-invariant features, while the multimodal interactive fusion further reinforces feature stability under complex capturing angles, thereby adapting to the practical application scenarios of multi-angle training and multi-angle competition recording in martial arts.

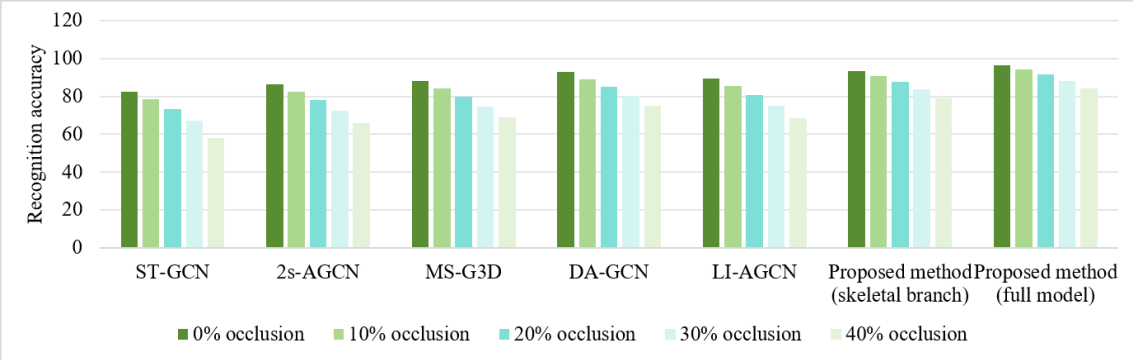


**Figure 9.** Recognition accuracy comparison under different occlusion rates (on the self-constructed martial arts dataset)
Note: ST-GCN = Spatial Temporal Graph Convolutional Network; 2s-AGCN = Two-Stream Adaptive Graph Convolutional Network; MS-G3D = Multi-Scale Graph 3D Convolutional Network; DA-GCN = Deformable Adaptive Graph Convolutional Network; LI-AGCN = Lightweight Initialization-Enhanced Adaptive Graph Convolutional Network

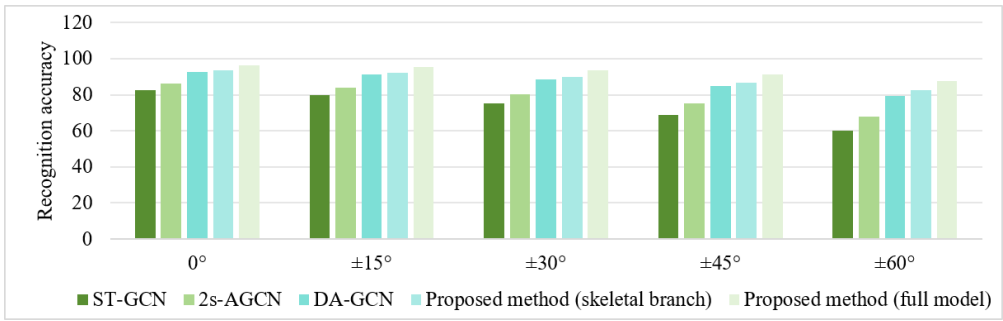


**Figure 10.** Recognition accuracy comparison under different viewpoint variations (on the self-constructed martial arts dataset)
Note: ST-GCN = Spatial Temporal Graph Convolutional Network; 2s-AGCN = Two-Stream Adaptive Graph Convolutional Network; DA-GCN = Deformable Adaptive Graph Convolutional Network

Overall, the systematic experimental results presented in this chapter comprehensively demonstrate that optimal overall performance is achieved by the proposed framework in both general-purpose action recognition and martial arts-specific action recognition tasks. Effective synergy is formed among all core modules, enabling lightweight real-time inference while maintaining ultra-high recognition accuracy, with excellent robustness exhibited under complex scenarios such as occlusion and viewpoint shifts, fully satisfying the engineering application requirements of intelligent martial arts training and real-time quantitative action analysis.

## 4. CONCLUSION

In response to the problems that existing skeleton-based action recognition models are inadequate for adapting to the specialized motion characteristics of martial arts—namely dynamic joint coordination and phased temporal evolution—and are incapable of simultaneously satisfying the requirements of high-precision recognition and real-time edge inference, an end-to-end real-time action recognition and intelligent analysis framework tailored for martial arts athletes was constructed in this study. Through the deformable adaptive graph convolution module, human anatomical priors were integrated with data-driven dynamic topology learning, by which the spatial modeling limitations of conventional fixed graph convolution were overcome, and the transient and variable limb coordination patterns of martial arts movements were precisely characterized. Through the martial-arts-semantics-driven multi-scale temporal modeling and phase perception mechanism, unsupervised adaptive segmentation of action temporal phases and differentiated scale feature extraction were accomplished, enabling effective mining of the hierarchical temporal rhythms and force-generation semantics of martial arts actions. Through the integration of grouped convolution, temporal sparse sampling, and knowledge distillation, a multi-level lightweight optimization system was constructed. With the introduction of an RGB-skeletal cross-attention fusion mechanism, computational overhead was compressed while the feature deficiencies of single modalities were compensated, significantly enhancing recognition robustness under complex scenarios. Systematic experiments on multiple benchmark datasets and a self-constructed specialized dataset demonstrated that excellent performance was achieved by the proposed method in both general-purpose action recognition and martial arts-specific recognition tasks. A cross-subject recognition accuracy of 93.9% was attained on NTU-60, and a recognition accuracy of 96.2% was achieved on the martial arts specialized dataset.

Stable real-time inference at 42.5 fps was realized on edge embedded devices, achieving an optimal multi-dimensional trade-off among recognition accuracy, inference efficiency, and environmental anti-interference capability.

The spatiotemporal feature modeling system for complex specialized sports is advanced through this study, and the application shortcomings of general-purpose graph convolutional networks in fine-grained martial arts action representation are addressed, providing reliable technical support for quantitative martial arts action evaluation, intelligent training guidance, and competitive performance analysis. The dynamic spatial adaptive modeling and phased temporal perception strategies proposed in this study can serve as general methodological references for various high-precision sports vision recognition tasks. In future work, the stability of the front-end pose estimation module will be further optimized to reduce the impact of keypoint detection noise on spatiotemporal modeling, while the application boundaries of the model will be extended to adapt the framework to intelligent analysis scenarios for multiple traditional sports and competitive activities. Additionally, through the integration of sports biomechanical parameters, the quantitative action analysis capability will be deepened, continuously promoting the deployment and development of intelligent vision technology in the field of sports science.

## REFERENCE


[1] Yue, R., Tian, Z., Du, S. (2022). Action recognition based on RGB and skeleton data sets: A survey. Neurocomputing, 512: 287-306. https://doi.org/10.1016/j.neucom.2022.09.071

[2] Xing, Y., Zhu, J. (2021). Deep learning-based action recognition with 3D skeleton: A survey. CAAI Transactions on Intelligence Technology, 6(1): 80-92. https://doi.org/10.1049/cit2.12014

[3] Feng, L., Zhao, Y., Zhao, W., Tang, J. (2022). A comparative review of graph convolutional networks for human skeleton-based action recognition. Artificial Intelligence Review, 55(5): 4275-4305. https://doi.org/10.1007/s10462-021-10107-y

[4] Xin, W., Liu, R., Liu, Y., Chen, Y., Yu, W., Miao, Q. (2023). Transformer for skeleton-based action recognition: A review of recent advances. Neurocomputing, 537: 164-186. https://doi.org/10.1016/j.neucom.2023.03.001

[5] Shi, L., Zhang, Y., Cheng, J., Lu, H. (2020). Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Transactions on Image Processing, 29: 9532-9545. https://doi.org/10.1109/TIP.2020.3028207

[6] Song, Y.F., Zhang, Z., Shan, C., Wang, L. (2020). Richly activated graph convolutional network for robust skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 31(5): 1915-1925. https://doi.org/10.1109/TCSVT.2020.3015051

[7] Lee, J., Jung, H. (2020). TUHAD: Taekwondo unit technique human action dataset with key frame-based CNN action recognition. Sensors, 20(17): 4871. https://doi.org/10.3390/s20174871

[8] Luo, C., Kim, S.W., Park, H.Y., Lim, K., Jung, H. (2023). Agnostic taekwondo action recognition using synthesized two-dimensional skeletal datasets. Sensors, 23(19): 8049. https://doi.org/10.3390/s23198049

[9] Lim, J., Luo, C., Lee, S., Song, Y.E., Jung, H. (2024). Action recognition of taekwondo unit actions using action images constructed with time-warped motion profiles. Sensors, 24(8): 2595. https://doi.org/10.3390/s24082595

[10] Cunha, P., Barbosa, P., Ferreira, F., et al. (2024). User assessment of a customized taekwondo athlete performance cyber–physical system. Applied Sciences, 14(11): 4683. https://doi.org/10.3390/app14114683

[11] Song, Y.F., Zhang, Z., Shan, C., Wang, L. (2023). Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2): 1474-1488. https://doi.org/10.1109/tpami.2022.3157033

[12] Liu, X., Li, Y., Xia, R. (2021). Adaptive multi-view graph convolutional networks for skeleton-based action recognition. Neurocomputing, 444: 288-300. https://doi.org/10.1016/j.neucom.2020.03.126

[13] Xie, J., Miao, Q., Liu, R., et al. (2021). Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition. Neurocomputing, 440: 230-239. https://doi.org/10.1016/j.neucom.2021.02.001

[14] Xiong, X., Min, W., Wang, Q., Zha, C. (2022). Human skeleton feature optimizer and adaptive structure enhancement graph convolution network for action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 33(1): 342-353. https://doi.org/10.1109/TCSVT.2022.3201186

[15] Huang, Z., Qin, Y., Lin, X., Liu, T., Feng, Z., Liu, Y. (2022). Motion-driven spatial and temporal adaptive high-resolution graph convolutional networks for skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 33(4): 1868-1883. https://doi.org/10.1109/TCSVT.2022.3217763

[16] Wu, L., Zhang, C., Zou, Y. (2023). Spatio temporal focus for skeleton-based action recognition. Pattern Recognition, 136: 109231. https://doi.org/10.1016/j.patcog.2022.109231

[17] Ding, C., Wen, S., Ding, W., Liu, K., Belyaev, E. (2022). Temporal segment graph convolutional networks for skeleton-based action recognition. Engineering Applications of Artificial Intelligence, 110: 104675. https://doi.org/10.1016/j.engappai.2022.104675

[18] Alsarhan, T., Ali, U., Lu, H. (2022). Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition. Computer Vision and Image Understanding, 216: 103348. https://doi.org/10.1016/j.cviu.2021.103348

[19] Xia, R., Li, Y., Luo, W. (2021). LAGA-Net: Local-and-global attention network for skeleton based action recognition. IEEE Transactions on Multimedia, 24: 2648-2661. https://doi.org/10.1109/TMM.2021.3086758

[20] Yu, L., Tian, L., Du, Q., Bhutto, J.A. (2022). Multi-stream adaptive spatial-temporal attention graph convolutional network for skeleton-based action recognition. IET Computer Vision, 16(2): 143-158. https://doi.org/10.1049/cvi2.12075

[21] Zhang, G., Wen, S., Li, J., Che, H. (2023). Fast 3D-graph convolutional networks for skeleton-based action recognition. Applied Soft Computing, 145: 110575.

https://doi.org/10.1016/j.asoc.2023.110575
[22] Dai, C., Lu, S., Liu, C., Guo, B. (2024). A light-weight skeleton human action recognition model with knowledge distillation for edge intelligent surveillance applications. Applied Soft Computing, 151: 111166. https://doi.org/10.1016/j.asoc.2023.111166
[23] Jiang, Y., Deng, H. (2024). Lighter and faster: A multi-scale adaptive graph convolutional network for skeleton-based action recognition. Engineering Applications of Artificial Intelligence, 132: 107957. https://doi.org/10.1016/j.engappai.2024.107957