

APPLYING C3D MODEL AND GRAPH ATTENTION NETWORK (GAT) TO SOLVING COMPLEX SCENE ACTIONS IN DANCE CHOREOGRAPHY

JINGYA CHEN¹, QIONG YANG² AND YANNI LIN^{3,*}

¹School of Education and Music
Hainan Vocational University of Science and Technology
No. 18, Qiongzhan Avenue, Meilan District, Haikou 571126, P. R. China
18876636800@163.com

²College of Information Technology

³Conservatory of Music
Qiongtai Normal University
No. 8, Zhongshan Road, Qiongzhan District, Haikou 571100, P. R. China
13976668796@163.com; *Corresponding author: linyanni@mail.qtnu.edu.cn

Received October 2025; revised February 2026

ABSTRACT. *In the actual scene of dance choreography, action recognition is often challenged by mutual occlusion between dancers and severe background interference, resulting in insufficient extraction of key features and reduced recognition accuracy. This paper uses C3D (Convolutional 3D) to capture high-level features with spatiotemporal continuity from the original video and effectively model the overall dynamic trend of dance actions; at the same time, OpenPose is used to extract the dancer's keypoint sequence and construct a dynamic human body relationship diagram with the dancer as the node; the GAT (Graph Attention Network) mechanism is applied to performing weighted fusion of feature information of neighboring nodes through the self-attention mechanism strategy, so that the model can still infer the complete action semantics with the help of contextual relationships when some keypoints are missing. Experimental results demonstrate that this method performs well in complex choreography scenarios, achieving an accuracy of $87.3 \pm 1.4\%$. On a test subset with multiple people and high occlusion, the model's performance degrades slightly (7.8%), demonstrating its effectiveness in dance action recognition under complex occlusion and background interference.*

Keywords: Choreography scene, Action recognition, Occlusion scene, Space-time continuity, Graph attention network

1. Introduction. With the in-depth application of intelligent visual technology in artistic performance analysis, dance action recognition has received widespread attention as a key supporting technology for human-computer interaction, virtual rehearsal assistance, and automatic scoring systems [1-3]. However, in real dance choreography scenes, problems such as frequent cross-movement of dancers, mutual occlusion of bodies, and complex and changing stage backgrounds seriously interfere with the integrity of visual features, resulting in instability in traditional action recognition methods in key gesture capture and semantic understanding [4,5]. Especially in group dances or highly synchronized choreography, individual actions are easily obscured by neighboring dancers, resulting in missing keypoints and broken spatiotemporal context, which in turn leads to misjudgment and missed detection [6-8]. Therefore, how to achieve robust and accurate action recognition

in high-occlusion and strong interference environments has become a core challenge in the current field of dance video analysis.

This paper focuses on fine-grained semantic parsing of actions in multi-person dance choreography scenarios. The research subjects cover a variety of forms, including modern dance, folk group dance, and ballet group dance, which are highly dynamic, collaborative, and spatially complex. Such scenarios require models to not only accurately identify the action categories of individual dancers, but also understand the formation transformation logic and action coordination relationships between the group. Patel-Grosz et al. [9] pointed out that dance not only has the characteristics of other basic human expressions, such as visual narrative and music, but also has similarities with the gestures of sign language and non-sign language speakers. Orgs et al. [10] found that the synchronization of actions between performers is the best indicator for predicting brain synchronization between expert and novice audiences. Qin et al. [11] proposed a spatiotemporal skeleton modeling framework that combined the part joint attention mechanism with a dynamic graph convolutional network to improve the capture of subtle and large-scale human movements. Zhang [12] designed an improved graph convolutional neural network algorithm for dance tracking and posture estimation to solve the problem of low dance action recognition accuracy caused by complex posture changes in dance actions. Sun and Song [13] proposed a posture estimation method based on an improved high-resolution network, with an accuracy of 73.5% and 79.5% on the MPII dataset and a self-built dance dataset, respectively. These studies show that action analysis for dance choreography must take account of both individual behavior perception and group context reasoning.

To address the challenge of action recognition in complex scenarios, researchers have proposed a variety of solutions. Sarma et al. [14] proposed a two-stream fusion model for gesture recognition. The two-stream network consists of two layers: a 3D convolutional neural network with gesture video as input and a 2D-CNN (Convolutional Neural Network) with motion template image guided by optical flow as input to improve the recognition effect. Xu et al. [15] proposed a two-stream human action recognition model based on human skeleton and scene image. The model fully utilized the advantages of skeleton information in motion expression and the advantages of image in scene presentation, and complementarily integrated scene information with human skeleton limbs based on spatiotemporal graph convolution to achieve human action recognition. Tian et al. [16] proposed a spatiotemporal graph convolutional neural network that integrated enhanced efficient channel attention mechanism. By establishing a dynamic mapping function between joint angular velocity and channel weight, it solved the bottleneck of insufficient spatiotemporal feature decoupling of traditional skeleton action recognition model in dance actions such as high-speed rotation. Qu [17] proposed a Transformer convolutional neural network model that integrated dynamic flow and static flow to conduct in-depth analysis and evaluation of dance actions. Wu et al. [18] constructed part-based graphs from different perspectives and employed a dual-branch framework to simultaneously capture static spatial and dynamic temporal representations, achieving a classification accuracy of 94.8%. In summary, existing methods either focus on local structure while ignoring overall movement trends or rely on ideal observation conditions, failing to effectively address the core bottlenecks of spatiotemporal feature fragmentation and context loss under high occlusion conditions.

This paper specifically explores how to achieve robust individual action unit recognition in complex group dances. On the one hand, the original video frame is input into the C3D network, and the global spatiotemporal motion features are extracted layer by layer through three-dimensional convolution; on the other hand, a dynamic human body relationship diagram is constructed based on the keypoint sequence of multiple dancers detected by OpenPose. Each dancer is used as a node, and adjacency connections are

established in real time based on the relative position in space. Confidence-weighted body center and displacement information are applied as node features. Message propagation is performed through a multi-layer GAT, and an adaptive attention mechanism is used to aggregate neighborhood information to enable semantic reasoning of actions for occluded individuals. The two-branch features are concatenated and fused at high levels, combining attention pooling with a fully connected classifier to output action categories. This method innovatively constructs a time-varying dynamic graph structure to respond to formation changes, designs a weak observation mechanism to maintain contextual integrity under occlusion, and significantly improves the robustness and accuracy of action recognition in high-density occlusion scenarios by leveraging the complementary two-stream approach.

To address the aforementioned challenges, this paper proposes a novel two-stream action recognition framework that integrates C3D and Graph Attention Networks (GAT) for complex dance choreography scenes. The main contributions of this work are threefold. 1) A dynamic human relationship graph constructed from multi-person keypoints is introduced, which evolves with spatial proximity to model group interactions in real time. 2) A weak observation mechanism combined with random node masking during training is designed, enabling the GAT branch to perform robust action inference even when keypoints are occluded or missing. 3) A multimodal fusion strategy is developed that concatenates global spatiotemporal features from C3D with context-aware graph features from GAT, significantly improving recognition accuracy under heavy occlusion and background interference. Compared to existing methods, the proposed approach explicitly leverages group-level contextual semantics to compensate for individual information loss, achieving superior robustness and precision in highly dynamic and crowded choreography scenarios.

2. Algorithm Design.

2.1. Overall framework design. This paper proposes a two-stream fusion action recognition framework for complex dance choreography scenarios. The system employs an end-to-end trainable architecture, taking as input a fixed-length video clip and its corresponding pose sequence, and outputting the class labels for the dominant dance actions within that time segment.

The framework consists of two parallel branches: a C3D-based spatiotemporal feature extraction pipeline that accepts raw RGB (Red, Green, Blue) video frames as input, and a graph-based structured reasoning pipeline based on a graph attention network that takes as input a sequence of keypoints extracted from multiple dancers using OpenPose. The two branches are strictly aligned in the temporal dimension to ensure temporal consistency in feature extraction. The C3D branch extracts spatiotemporal features step by step through convolution and outputs a global motion representation after global pooling. The GAT branch treats each dancer as a node in the graph and constructs a per-frame human relationship graph based on their keypoint trajectories, which evolves over time to form a dynamic graph sequence. The graph structure dynamically updates edge connections based on the real-time spatial distances between dancers, supporting time-varying modeling of adjacency relationships. The features extracted from the two branches are concatenated and fused at a high level: the C3D output reflects overall body displacement trends and rhythmic changes, while the GAT output encodes relative positions, collaborative behaviors, and contextual dependencies between individuals. The concatenated joint features are fed into the classification module to achieve multimodal semantic complementarity. All components share a unified time index, and the video frames and posture detection results have completed frame-level synchronization calibration in the preprocessing stage. The overall framework is shown in Figure 1.

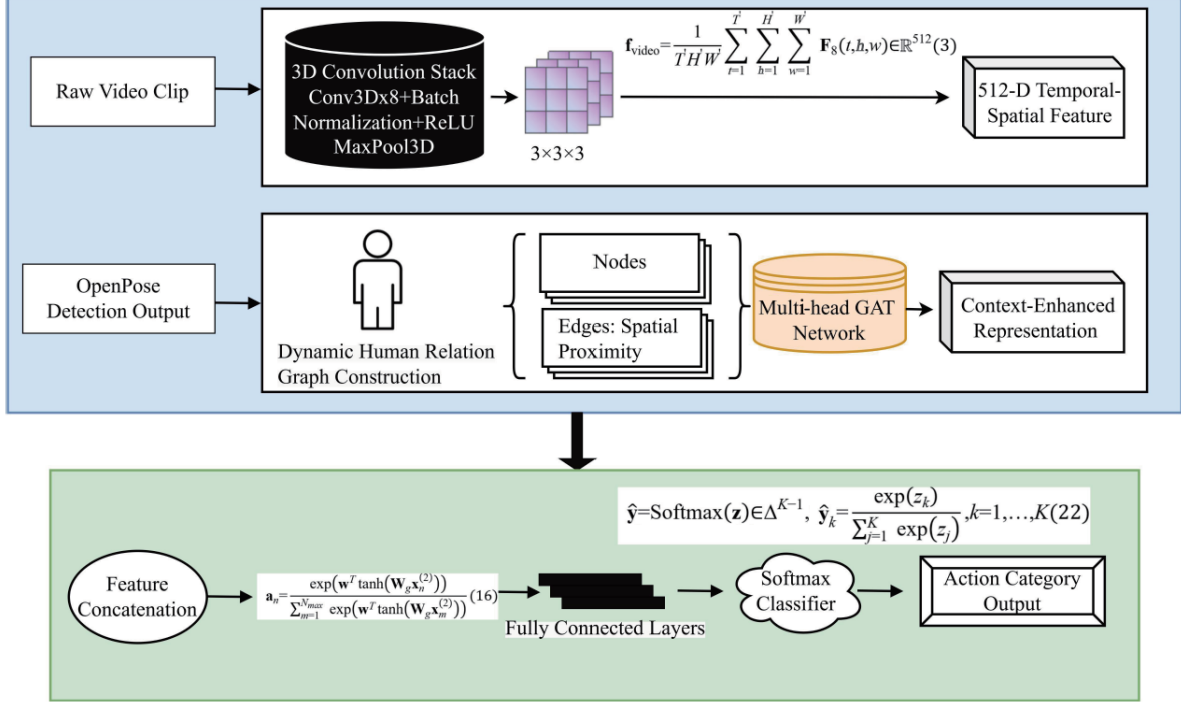


FIGURE 1. Overall framework diagram

Figure 1 shows the overall process of the C3D-GAT dual-stream fusion framework proposed in this paper: one branch inputs the original video clip into the C3D network, extracts spatiotemporal features layer by layer through three-dimensional convolution, and finally outputs a 512-dimensional global motion representation; the other branch constructs a dynamic human body relationship graph with dancers as nodes and spatial proximity as edges based on the keypoint sequence of multiple dancers detected by OpenPose, and feeds it into the graph attention network for message propagation. The adaptive attention mechanism is used to aggregate neighbor node information to generate context-enhanced structured features; after the two branches are strictly aligned in the temporal dimension, the visual features extracted by C3D are spliced and fused with the graph features output by GAT, and the final action category is output through the fully connected layer and Softmax classifier to achieve robust recognition of complex dance choreography actions.

2.2. Dual-stream input feature extraction. The video stream takes a continuous RGB video clip $V \in \mathbb{R}^{T \times H \times W \times 3}$ as input, where T represents the number of temporal frames, and H and W represent the spatial height and width, respectively. The clip is first normalized to the interval $[0, 1]$, and the ImageNet mean is subtracted before being fed into the C3D network. The C3D backbone consists of 8 three-dimensional convolutional layers $\{\text{Conv3D}_k\}_{k=1}^8$, each layer using a $3 \times 3 \times 3$ convolution kernel with the “same” padding to preserve the spatiotemporal dimensions [19,20]. The output of the k -th layer is defined as

$$\mathbf{F}_k = \text{ReLU}(\text{BN}(\text{Conv3D}_k(\mathbf{F}_{k-1}))), \quad \mathbf{F}_0 = V \quad (1)$$

In Formula (1), $\text{BN}(\cdot)$ represents the batch normalization operation. Every two convolutional layers are followed by a $2 \times 2 \times 2$ maximum pooling operation [21]:

$$\mathbf{F}_k = \max_{\text{pool}}^{(2,2,2)}(\mathbf{F}_{k-1}), \quad k \in \{2, 4, 6, 8\} \quad (2)$$

Finally, the compact spatiotemporal feature vector is obtained through global average pooling [22]:

$$\mathbf{f}_{\text{video}} = \frac{1}{T'H'W'} \sum_{t=1}^{T'} \sum_{h=1}^{H'} \sum_{w=1}^{W'} \mathbf{F}_8(t, h, w) \in \mathbb{R}^{512} \quad (3)$$

The pose stream is constructed based on the OpenPose v1.7.0 detection results [23,24]. Let the set of body keypoints of the n -th dancer in the t -th frame be $s_{n,t} \in \mathbb{R}^{18 \times 2}$, which contains the two-dimensional coordinates (x, y) of 18 joint points. The complete sequence is stacked in time to form the original pose tensor $P \in \mathbb{R}^{T \times N \times 18 \times 2}$, where N is the number of dancers that can be detected in the current frame. In order to achieve fixed-dimensional input, a zero-filling mechanism is applied: the maximum number of supported people is set to $N_{\max}$. If the number of dancers in the current frame is $N < N_{\max}$, then zero vectors are filled in the $N + 1$ to $N_{\max}$ position; if $N > N_{\max}$, only the pose data of the first $N_{\max}$ dancer is retained. The tensor after zero filling is recorded as $\hat{P} \in \mathbb{R}^{T \times N_{\max} \times 18 \times 2}$, and Z-score normalization is further performed on the coordinates of each keypoint [25]:

$$\hat{p}_{n,t,i,j} = \frac{p_{n,t,i,j} - \mu_j}{\sigma_j + \epsilon}, \quad j \in \{x, y\} \quad (4)$$

In Formula (4), μ_j and σ_j are the mean and standard deviation of all x or y coordinates in the training set; i is the joint index; j is the coordinate axis direction; ϵ is the smoothing term. The normalized pose sequence serves as the initial node feature input for the GAT branch. The two data streams are strictly aligned on the time axis: the video clips and pose detection use the same start and end frame indices and a unified sampling rate. All operations are frame-level aligned by synchronizing timestamps during the data loading phase.

2.3. Modeling of dynamic human body relationship graph. In order to explicitly depict the spatial coordination and interactive evolution relationship between dancers in dance choreography, this paper constructs a time-varying graph structure and models the multi-dancer system in each frame of video as a dynamic undirected graph. Each video frame t is modeled as an independent human body relationship graph $G_t = (V, E_t)$ to achieve a dynamic representation of the group behavior topology [26]. The node set $V = \{v_n\}_{n=1}^{N_{\max}}$ in the graph corresponds to the maximum number of $N_{\max}$ dancers in the scene, and each node v_n represents the motion state of the n -th dancer. Its initial node feature vector $\mathbf{x}_n^{(0)} \in \mathbb{R}^4$ is composed of the weighted body center position and two-dimensional displacement:

$$\mathbf{x}_n^{(0)} = [c_{n,t}^x, c_{n,t}^y, \Delta x_n, \Delta y_n]^T \quad (5)$$

The body center coordinate $(c_{n,t}^x, c_{n,t}^y)$ is calculated based on the weighted confidence of keypoints [27]:

$$c_{n,t}^x = \frac{\sum_{k=1}^K s_{n,t,k} \cdot p_{n,t,k}^x}{\sum_{k=1}^K s_{n,t,k}}, \quad c_{n,t}^y = \frac{\sum_{k=1}^K s_{n,t,k} \cdot p_{n,t,k}^y}{\sum_{k=1}^K s_{n,t,k}} \quad (6)$$

In Formula (6), K is the number of joint points; $p_{n,t,k}^x$ and $p_{n,t,k}^y$ are the coordinates of the k -th joint; $s_{n,t,k} \in [0, 1]$ is the detection confidence. The displacement component is defined as the change in the center position between adjacent frames:

$$\Delta x_n = c_{n,t}^x - c_{n,t-1}^x, \quad \Delta y_n = c_{n,t}^y - c_{n,t-1}^y \quad (7)$$

For the first frame, assuming $\Delta x_n = \Delta y_n = 0$, the edge set $E_t \subseteq V \times V$ is dynamically constructed based on spatial proximity. An undirected edge exists between any two nodes

if and only if their Euclidean distance is below a threshold. The corresponding edge weight $w_{ij}^{(t)}$ is calculated using the Gaussian kernel function [28,29]:

$$w_{ij}^{(t)} = \exp \left(-\frac{\|\mathbf{c}_{i,t} - \mathbf{c}_{j,t}\|_2^2}{2\sigma^2} \right) \quad (8)$$

In Formula (8), $\mathbf{c}_{i,t}$ represents the coordinates of the body center of the i -th dancer in the t -th frame, and $\mathbf{c}_{j,t}$ represents the coordinates of the body center of the j -th dancer in the t -th frame. This weight reflects the interaction strength between dancers and is used for neighborhood aggregation in the subsequent graph attention mechanism. To address the problem of missing keypoints caused by occlusion, a “weak observation” mechanism is applied: if the average keypoint confidence of a node v_n is $\bar{s}_{n,t} = \frac{1}{K} \sum_{k=1}^K s_{n,t,k} < 0.3$, it is marked as a weak observation node. At this time, the estimated position of the previous moment is retained as the current position input and multiplied by the attenuation factor before the initial feature:

$$\tilde{\mathbf{x}}_n^{(0)} = \gamma \cdot \mathbf{x}_n^{(0)} \quad (9)$$

At the same time, it is incorporated into the graph structure to participate in message passing, ensuring that even if individual information degrades, the integrity of the group context can still be maintained through graph connections. Ultimately, the serialized graph collection $\{G_t\}_{t=1}^T$ constitutes a dynamic human relationship graph stream, which serves as the input of the graph attention network and supports fine-grained modeling of complex choreography formation evolution and collaborative actions.

2.4. Contextual reasoning with graph attention mechanism. In order to achieve robust inference of individual actions in complex dance choreography scenes, this paper applies a multi-layer graph attention network based on the dynamic human body relationship graph. By adaptively weighting and aggregating neighborhood information, it models the non-local interaction dependencies between dancers and maintains semantic consistency under the condition of partial loss of keypoints [30]. For the h -th attention head in the layer l , the node features are first linearly transformed:

$$\mathbf{z}_n^{(l,h)} = \mathbf{W}_h^{(l)} \mathbf{x}_n^{(l-1)} \in \mathbb{R}^{d_h} \quad (10)$$

Then, the unnormalized attention coefficient of the node i to its neighbor $j \in N(i)$ is calculated:

$$e_{ij}^{(l,h)} = \text{LeakyReLU} \left(\mathbf{a}_h^{(l)T} \left[\mathbf{z}_i^{(l,h)} \parallel \mathbf{z}_j^{(l,h)} \right] \right) \quad (11)$$

In Formula (11), $\parallel$ represents the vector concatenation operation; $\mathbf{a}_h^{(l)}$ is a learnable attention weight vector. The attention weight is normalized within the neighborhood by the softmax function [31]:

$$\alpha_{ij}^{(l,h)} = \frac{\exp \left(e_{ij}^{(l,h)} \right)}{\sum_{k \in N(i)} \exp \left(e_{ik}^{(l,h)} \right)} \quad (12)$$

Each attention head aggregates neighborhood information independently:

$$\tilde{\mathbf{x}}_i^{(l,h)} = \sum_{j \in N(i)} \alpha_{ij}^{(l,h)} \mathbf{z}_j^{(l,h)} \quad (13)$$

The outputs of all heads are concatenated in the feature dimension to form the final output of the current layer:

$$\mathbf{x}_i^{(l)} = \sigma \left(\parallel_{h=1}^H \tilde{\mathbf{x}}_i^{(l,h)} \right) \quad (14)$$

In Formula (14), $\sigma(\cdot) = \text{ReLU}$. The second layer GAT takes the output of the first layer as input and adopts the same structure but uses average merging instead of splicing to control the dimension growth:

$$\mathbf{x}_i^{(2)} = \sigma \left(\frac{1}{H} \sum_{h=1}^H \tilde{\mathbf{x}}_i^{(2,h)} \right) \quad (15)$$

To further improve the model's generalization capabilities in occluded scenarios, a random node masking strategy is applied during training: with probability $p \sim U(0.1, 0.4)$, initial features $\mathbf{x}_n^{(0)}$ are set to zero or superimposed with Gaussian noise $\epsilon_n \sim N(0, 0.1^2 I)$ to simulate varying degrees of keypoint loss. This mechanism forces the GAT to rely on contextual information to complete state reconstruction within the constraints of the network topology, enabling implicit reasoning about the semantics of the occluded dancer's actions.

Ultimately, the graph attention outputs $\{\mathbf{x}_n^{(2)}\}_{n=1}^{N_{\max}}$ of all frames form a structure-aware sequence of action representations, which are fed into the subsequent feature fusion module. The attention weights are shown in Figure 2. There is no explicit decoder network or dedicated reconstruction loss function during training. The model learns to generate context-enriched node features through the message-passing mechanism of GAT, driven solely by the final cross-entropy loss of the classification task. Any assessment of keypoint recovery is performed post-hoc for evaluation purposes only, without influencing the training process.

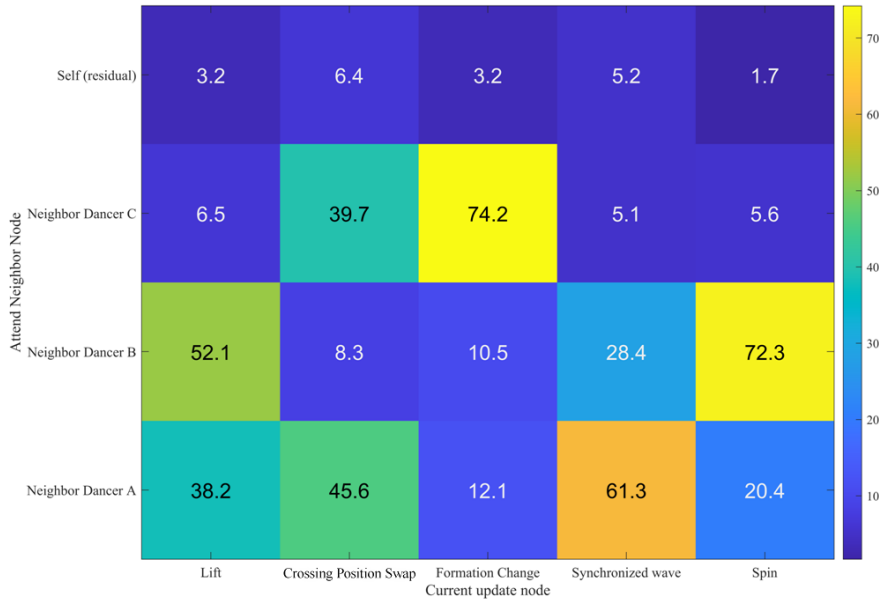


FIGURE 2. Attention weight heat map

Figure 2 shows the distribution of attention weights learned by the GAT module in a dance choreography scenario. The horizontal axis represents the dancer's action currently being updated, and the vertical axis represents the neighboring nodes it focuses on (including other dancers and its own residuals). Attention among dancers is not evenly distributed, but rather dynamically focuses on key collaborators based on the semantics of the actions. Dancers performing a lifting action assign a high weight of 52.1% to neighbor Dancer B. Crossing Position Swap relies on the synchronization of Dancers A and C's actions to determine path interactions, while dancers performing a formation change pay close attention to neighbor Dancer C (74.2%). This demonstrates that the model captures the functional role relationships and interaction priorities in complex choreography

through an adaptive attention mechanism, enabling contextual awareness and action semantic reasoning in the presence of occlusion or information loss.

The effectiveness of the GAT branch in handling occlusion stems from two synergistic mechanisms: the weak observation mechanism and random node masking during training. The weak observation mechanism ensures that even when a dancer's keypoint confidence is low, the node remains active in the graph and its position is softly updated using the previous state. This prevents the abrupt removal of a dancer from the group context, forcing the model to learn to reason about individuals whose signals are degraded but not absent. More critically, the random node masking strategy acts as a powerful regularizer. By randomly setting a node's initial features to zero or adding noise during training, we simulate severe, unpredictable occlusions. This forces the GAT to rely less on the raw input of any single node and more on the topological relationships and collective dynamics encoded in the graph structure.

To address the instability caused by rapid position changes in dance choreography, where frequent edge activation and deactivation may lead to graph structure jitter, an adaptive threshold determination mechanism is designed. Instead of using a fixed pixel distance, the spatial proximity threshold is dynamically calibrated based on the normalized dancer scale within each video frame. Specifically, the median torso length of all detected dancers in the current frame is computed as a reference scale, and the connection threshold is set as a multiple of this scale (empirically set to 1.5 times the median torso length). This adaptive strategy ensures that the graph structure remains relatively stable across different shooting distances and dancer sizes, reducing spurious edge fluctuations caused by camera perspective changes or individual scale variations.

Furthermore, a temporal smoothing mechanism is applied to the edge activation process. An edge is considered valid only if the spatial proximity condition is maintained consecutively for at least two frames, effectively filtering out transient connections caused by momentary cross-overs or pose estimation jitter. For formation transformations where dancers gradually move apart, the Gaussian kernel weighting in Equation (8) naturally attenuates edge influence as distance increases, rather than employing a hard cut-off threshold. This soft weighting mechanism allows the graph to maintain weak but meaningful connections during formation transitions, preserving contextual information without introducing abrupt topological changes. Experimental observations confirm that this adaptive thresholding approach maintains stable graph structures across diverse choreography patterns, including rapid formation changes in modern dance and synchronized group movements in folk dance, demonstrating its generalizability across different dance styles and spatial configurations.

2.5. Multimodal feature fusion and action classification. Based on the global motion representation extracted by the C3D branch and the context-enhanced graph features output by the GAT branch, this section designs a compact and differentiable multimodal fusion mechanism and constructs an end-to-end classification network to achieve accurate semantic discrimination of complex dance actions. To further extract the consensus representation of group actions, an attention-based pooling operation is used to aggregate all valid nodes [32,33]:

$$\mathbf{a}_n = \frac{\exp\left(\mathbf{w}^T \tanh\left(\mathbf{W}_g \mathbf{x}_n^{(2)}\right)\right)}{\sum_{m=1}^{N_{\max}} \exp\left(\mathbf{w}^T \tanh\left(\mathbf{W}_g \mathbf{x}_m^{(2)}\right)\right)} \quad (16)$$

In Formula (16), $\mathbf{W}_g$ and $\mathbf{w}$ are learnable parameters used to calculate the contribution weight of each dancer to the current action category. After weighting, the group feature

vector is obtained:

$$\mathbf{f}_{\text{pose}} = \sum_{n=1}^{N_{\max}} \mathbf{a}_n \cdot \mathbf{x}_n^{(2)} \quad (17)$$

This strategy avoids semantic dilution caused by simple averaging and highlights the influence of the dominant dancer or key interaction on the classification decision. $\mathbf{f}_{\text{video}}$ and $\mathbf{f}_{\text{pose}}$ are then concatenated with the data across modalities to form a joint feature. A three-layer fully connected network is applied to implementing classification mapping. The specific structure is as follows [34,35]:

First layer:

$$\mathbf{h}_1 = \text{ReLU}(\mathbf{W}_1 \mathbf{f}_{\text{fuse}} + \mathbf{b}_1) \quad (18)$$

Second layer (including regularization):

$$\mathbf{h}_2 = \text{ReLU}(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2) \quad (19)$$

Dropout operation is applied:

$$\tilde{\mathbf{h}}_2 = \text{Dropout}(\mathbf{h}_2; p = 0.5) \quad (20)$$

Third layer (output layer):

$$\mathbf{z} = \mathbf{W}_3 \tilde{\mathbf{h}}_2 + \mathbf{b}_3 \quad (21)$$

Finally, the normalized category probability distribution is generated by the Softmax function:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{z}) \in \Delta^{K-1}, \quad \hat{y}_k = \frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)}, \quad k = 1, \dots, K \quad (22)$$

The model is optimized with standard cross entropy loss:

$$\mathcal{L}_{\text{cls}} = - \sum_{k=1}^K \mathbf{y}_k \log \hat{\mathbf{y}}_k \quad (23)$$

All parameters are jointly updated through backpropagation, using the Adam optimizer to perform gradient descent $\theta \leftarrow \theta - \eta \cdot \tilde{\nabla}_{\theta} \mathcal{L}_{\text{cls}}$, where θ represents the complete set of learnable parameters, the experimental setting parameters are shown in Table 1. The entire C3D-GAT framework is trained end-to-end in a fully supervised manner, using only the classification cross-entropy loss; no additional self-supervised pre-training or auxiliary

TABLE 1. Parameter setting table

Hyperparameter	Value
Maximum number of dancers	8
Spatial distance threshold	Adaptive, based on normalized dancer height (reference: 1.5 m)
Gaussian kernel standard deviation	0.75 (scaled with distance threshold)
Weak observation attenuation factor	0.9
Number of GAT layers	2
Number of attention heads (L1)	8
Number of attention heads (L2)	4
Hidden dimension per head	64
Attention pooling output dimension	512
Feature concatenation	$[512] + [512] \rightarrow [1024]$
Fully connected layer 1	$1024 \rightarrow 512$
Fully connected layer 2	$512 \rightarrow 256$

reconstruction losses are employed. During the inference phase, the system processes the continuous video stream in a sliding window manner, independently completing forward propagation for each segment and outputting the action category corresponding to the maximum probability:

$$\hat{c} = \arg \max_k \hat{\mathbf{y}}_k \quad (24)$$

To ensure scale invariance, the spatial distance threshold and Gaussian kernel parameters are defined in normalized pixel space. The 1.5 meters and $\sigma = 0.75$ values in Table 1 are reference values. In actual calculations, adaptive pixel transformation is performed based on the median length of the dancer’s torso in the video to ensure robustness to different shooting distances and resolutions.

3. Experimentation and Verification.

3.1. Experimental design. The experiment is sampled from a number of professional performances and rehearsal scenes, including five typical group dance movements: lift, spin, formation change, synchronized wave, and crossing position swap. The training, validation, and test sets are split in a 7 : 1 : 2 ratio to ensure a balanced distribution of samples across categories. All models are trained on an NVIDIA Tesla V100 GPU (Graphics Processing Unit). The input is uniformly 32-frame, 224×224 video blocks, with pose sequences extracted using OpenPose. The same data preprocessing and training strategies are used throughout the experiment. The experimental setup is shown in Table 2. To ensure the statistical rigor and reproducibility of the reported results, all experiments were repeated five times. Random Seeds: Each repetition was conducted with a different random seed (seed = 0, 1, 2, 3, 4) to initialize the network weights and data shuffling, ensuring the variability captured is due to stochastic training dynamics rather than a single initialization artifact. Significance Testing: The statistical significance of performance differences between the proposed method and each baseline was assessed using a paired

TABLE 2. Experimental settings

Item	Configuration
Number of action classes	Lift, Spin, Formation Change, Synchronized Wave, Crossing Position Swap
Dataset size	1,248 choreography segments
Dancer count	2-8 dancers per scene
Environment types	Studio Rehearsals (40%), Stage Performances (50%), Outdoor Scenes (10%)
Splitting strategy	Cross-session and cross-performer
Data set	Training: 70%, Validation: 10%, Test: 20%
Video resolution	1920 × 1080 (original), Input size: 224 × 224
Frame rate	25 fps
Temporal window length	32 frames
Pose sequence completion	Missing keypoints filled via linear interpolation with confidence-weighted smoothing
Training platform	Single NVIDIA Tesla V100 GPU, PyTorch 1.12
Optimizer	Adam
Initial learning rate	1×10^{-4}
Batch size	16
Training epochs	60 (Early stopping: terminate if validation loss does not decrease for 3 consecutive epochs)

t-test. To ensure fair comparison, all baseline models were compared under unified experimental conditions. ResNet3D and I3D used pre-trained weights from Kinetics-400 and were fine-tuned on this dataset; ST-GCN was trained from scratch as a skeleton model. All methods used the same input data source and employed the same data augmentation strategy (random temporal cropping, horizontal flipping, and color jittering) and training configuration. This unified protocol ensured that performance differences were not due to the model structure itself.

3.2. Dance action recognition accuracy. To measure the model's ability to discriminate on the entire test set, this paper first uses recognition accuracy as the basic evaluation metric. The proportion of correctly classified samples in the test set to the total number of samples is calculated:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N I(\mathbf{y}_i = \hat{\mathbf{y}}_i) \quad (25)$$

In Formula (25), $\mathbf{y}_i$ is the true value label; $\hat{\mathbf{y}}_i$ is the model prediction result; $I(\cdot)$ is the indicator function.

Figure 3 compares the recognition accuracy of the proposed method, ResNet3D (Residual Network 3D), I3D (Inflated 3D ConvNet), and ST-GCN (Spatial-Temporal Graph Convolutional Network) on five typical dance actions. The horizontal axis represents Spin, Lift, Formation Change, Synchronized Wave, Crossing Position Swap, while the vertical axis represents recognition accuracy. The proposed method significantly outperforms ResNet3D, I3D, and ST-GCN in all action categories, achieving an average accuracy of $87.3 \pm 1.4\%$, a 7.2% improvement over ST-GCN. When dealing with tasks such as “formation change” that are highly dependent on the spatial relationship of the group, the proposed method achieves a recognition accuracy of $87.6 \pm 1.4\%$, showing strong adaptability to complex choreography patterns. In the low-occlusion action of “synchronized wave”, the proposed method maintains a leading position with $87.9 \pm 1.2\%$. The performance fluctuation of the proposed method in various actions is minimal, indicating that it has not only high accuracy but also better cross-task stability.

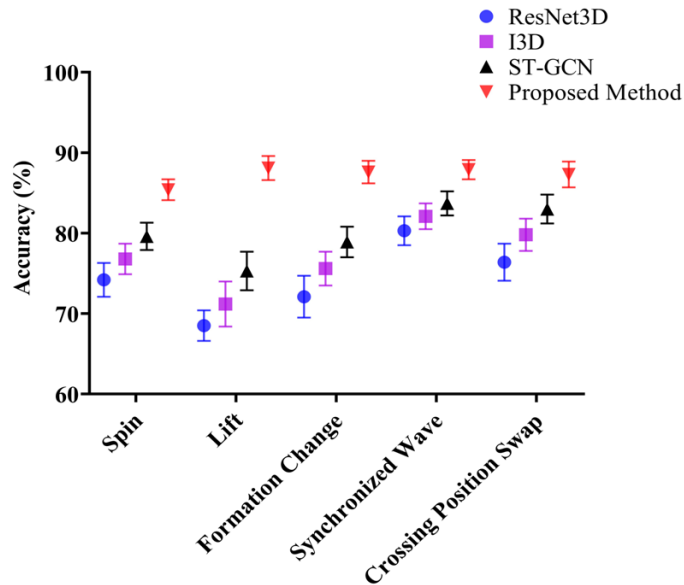


FIGURE 3. Recognition accuracy

The C3D branch effectively extracts the overall motion texture of actions such as “Spin” and “Synchronized Wave”, and has good background suppression capabilities; the GAT branch constructs a dynamic human body relationship graph and uses the attention mechanism to adaptively weigh the motion states of neighboring dancers in high-occlusion scenarios such as “Lift” and “Crossing Position Swap” to achieve contextual reasoning of the semantics of the occluded individual movements. When a dancer is completely occluded, the model can infer its movement intention through the relative position changes and attention weight distribution of the surrounding dancers. The dynamic edge weight update mechanism enables the graph structure to respond to the topological evolution in the “Formation Change” in real time, avoiding the modeling deviation caused by the fixed adjacency matrix.

3.3. Stability of highly occluded subset recognition. To specifically evaluate the model’s robustness under extreme visual degradation conditions, a high-occlusion subset is defined based on a quantifiable metric. The occlusion ratio for each choreography segment is calculated as the average percentage of a dancer’s 18 body joints that are either completely missing or have a detection confidence score below 0.3 across all frames within the clip. From the test set, a subset with an occlusion ratio greater than 40% is selected, defining this as the high-occlusion group. The accuracy on this subset is calculated and compared with the accuracy Acc_{full} on the full set:

$$\Delta Acc = \frac{Acc_{full} - Acc_{high-occ}}{Acc_{full}} \quad (26)$$

This indicator is used to quantify the degree of degradation of model performance in severe occlusion scenarios. The smaller the value, the stronger the stability and anti-interference ability of the model.

Figure 4 shows a comparison of the recognition accuracy of four methods on the full dataset and the high-occlusion subset, along with the corresponding accuracy degradation (ΔAcc). The horizontal axis shows ResNet3D, I3D, ST-GCN, and the proposed method. The proposed method achieves 87.3% on the full dataset and maintains 80.5% recognition accuracy under high-occlusion conditions, significantly exceeding the other methods.

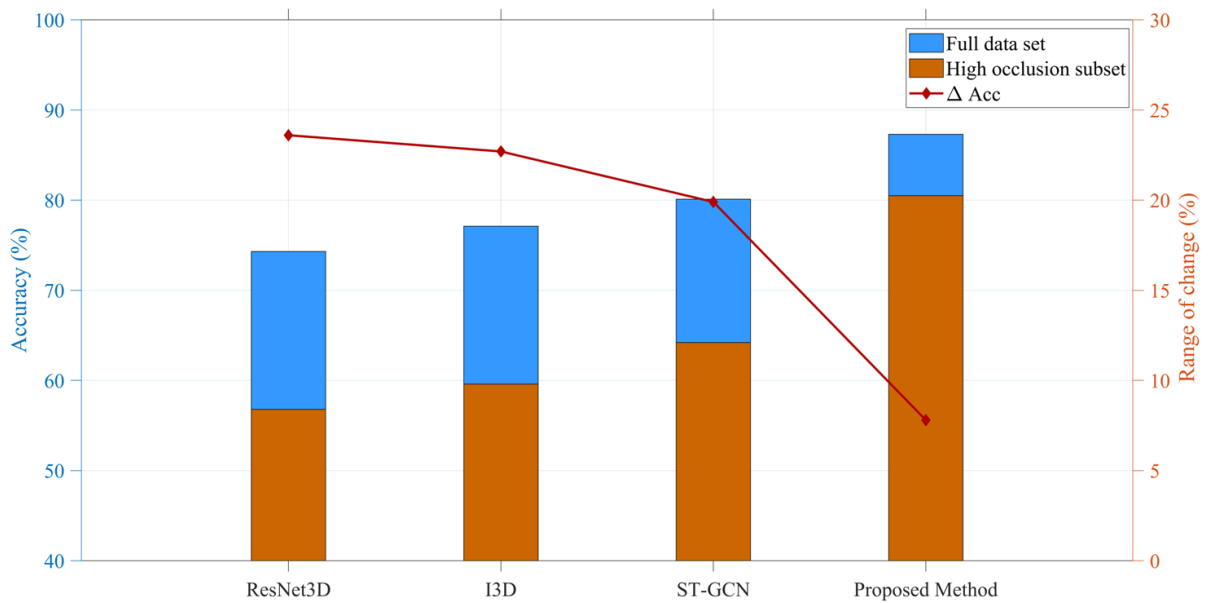


FIGURE 4. Recognition stability

ResNet3D drops from 74.3% to 56.8%, I3D from 77.1% to 59.6%, and ST-GCN from 80.1% to 64.2%, all showing significant declines. Meanwhile, the accuracy degradation of the proposed method is only 7.8%, far lower than the 23.6% of ResNet3D, the 22.7% of I3D, and the 19.9% of ST-GCN, highlighting its stable performance under extreme occlusion scenarios.

In the “Crossing Position Swap” scenario, when a dancer’s posture information is interrupted due to a brief occlusion, the model does not rely solely on linear extrapolation of the time series for prediction. Instead, it combines the mutation characteristics of the movement states of the surrounding dancers to identify potential interaction signals and accordingly enhance the influence weights of these key collaborators in the graph network, and reversely infer the possible movement trajectory of the individual and its functional role in the choreography. Dual-stream fusion is not a simple feature splicing. It facilitates the mutual verification and correction of visual motion cues and topological relationship reasoning at the semantic level, enabling the system to have stronger motion boundary discrimination capabilities when facing high-density interference, and significantly reducing the risk of misjudgment caused by modal deviation or perceptual blind spots.

3.4. Comprehensive evaluation of classification performance based on F1 score.

To further comprehensively evaluate the model’s ability to discriminate various dance actions, this paper uses the F1 score as the core comprehensive evaluation metric for classification performance. Precision $Prec_k$ and recall Rec_k are calculated for each action category, and the F1 score is then calculated:

$$F1_k = \frac{2 \cdot Prec_k \cdot Rec_k}{Prec_k + Rec_k} \quad (27)$$

Figure 5 systematically compares the F1 scores of the proposed method, ResNet3D, I3D, and ST-GCN on five typical dance action categories. The proposed method significantly outperforms ResNet3D, I3D, and ST-GCN in all categories. In the “Lift” action, ResNet3D’s F1 score is only $67.8 \pm 3.1\%$, reflecting its limited performance in complex interactive scenarios. The proposed method achieves an F1 score of $87.6 \pm 1.4\%$ for this action, significantly outperforming other methods. The overall F1 score of each method is $72.8 \pm 2.4\%$, $76.1 \pm 2.0\%$, and $79.2 \pm 1.7\%$, respectively, while the proposed method achieves a high F1 score of $86.7 \pm 1.3\%$. The inter-class variability is minimal, with the lowest standard deviation, demonstrating stable output across various action types.

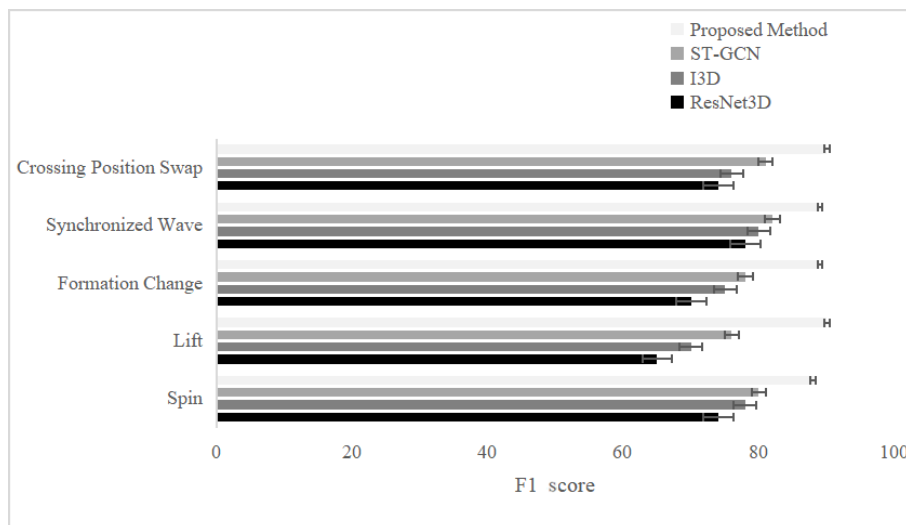


FIGURE 5. F1 score

ResNet3D relies on local appearance features and is prone to misjudgment when limbs overlap, resulting in low F1. Although ST-GCN is based on skeleton modeling, its fixed graph structure makes it difficult to adapt to dynamic formation changes, limiting F1 performance. The method in this paper constructs a dynamic human body relationship graph and applies a graph attention mechanism, which can use group context for semantic completion when individuals are occluded, significantly improving the recall rate. The global motion consistency prior provided by the C3D branch effectively suppresses erroneous predictions and ensures high accuracy. This dual-stream collaborative mechanism of “structure guidance + semantic completion” achieves a good balance between precision and recall, which is ultimately reflected in the F1 indicator.

3.5. Inter-class confusion degree. To quantitatively evaluate the model’s ability to distinguish similar dance moves, a normalized confusion matrix is constructed. All off-diagonal elements are extracted to form a set $E_{\text{off-diag}}$, and their standard deviation is calculated:

$$\sigma_{\text{conf}} = \sqrt{\frac{1}{|E|} \sum_{e \in E_{\text{off-diag}}} (e - \mu)^2}, \quad \mu = \frac{1}{|E|} \sum e \quad (28)$$

The standard deviation is used to measure the degree of dispersion of the overall confusion distribution of the model: the smaller the value, the lower the misclassification rate among categories and the more evenly distributed it is, indicating that the model has clear classification boundaries and stronger discrimination power; otherwise, it indicates that the model has serious and unbalanced category confusion problems.

Figure 6 shows the normalized confusion matrices for the proposed method, ResNet3D, I3D, and ST-GCN for five dance action categories. The horizontal column represents the predicted labels, and the vertical column represents the true labels. The value in each cell represents the percentage of that category being identified as another category. The overall distribution shows that ResNet3D exhibits significant misclassifications. In the “Lift” action, only 68.5% of samples are correctly identified, while 12.4% are misclassified as “Formation Change” and 9.8% as “Synchronized Wave”. This indicates that it struggles to distinguish semantically similar actions in complex interactive scenarios. While I3D and ST-GCN have shown some improvement, they still generally miss or misclassify over 4% of highly occluded actions, such as “Crossing Position Swap”. The proposed method shows higher accuracy, such as 86.7% for “Spin”, 88.1% for “Lift”, and 87.0% for “Crossing Position Swap”. The colors of non-diagonal areas are generally darker, with values mostly

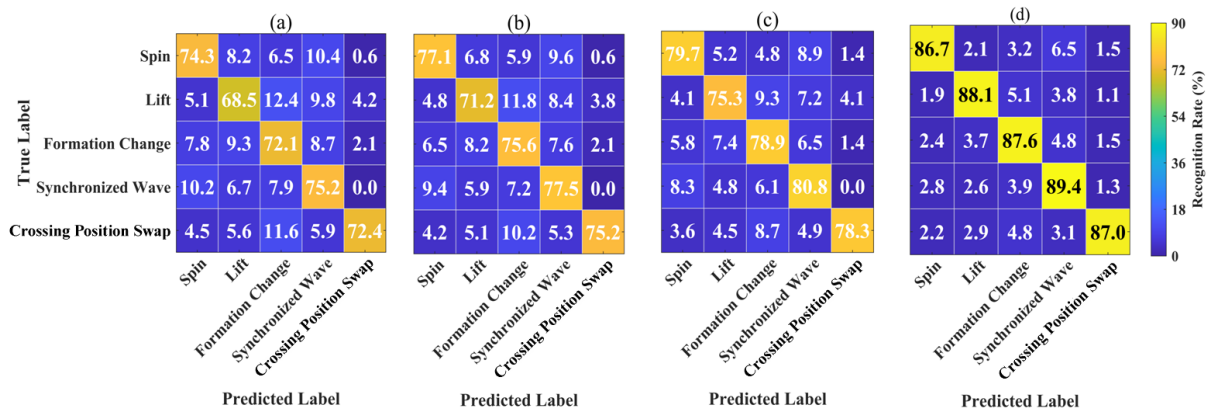


FIGURE 6. Inter-class confusion degree ($n = 250$): (a) ResNet3D; (b) I3D; (c) ST-GCN; (d) the method of this paper

below 3.9%, indicating that its misjudgment rate is significantly reduced, and the degree of confusion between categories is minimized.

Traditional methods rely only on local appearance or fixed skeleton topology, and can easily confuse different actions with similar motion patterns. For example, “Synchronized Wave” and “Formation Change” may converge in global optical flow, resulting in more than 7% cross-misjudgment in ResNet3D. The method in this paper constructs a dynamic human body relationship graph through GAT, and adaptively focuses on the relative position changes and collaboration timing between key dancers during the reasoning process, effectively distinguishing the essential differences between actions: even if the overall movement trends of two actions are similar, as long as the interaction patterns between individuals are different, the model can distinguish them through the differentiated distribution of attention weights. At the same time, the macro-rhythm information provided by the C3D branch further assists classification decisions, allowing the model to not only “see who is moving” but also “understand why they are moving in this way”, and ultimately establish clear classification boundaries between highly similar choreography actions.

3.6. Missing keypoint recovery success rate. To evaluate the model’s contextual reasoning capabilities, we fitted a ridge regression decoder ($\lambda = 1e-3$) to the trained GAT node features on an independent validation set to linearly map the features back to the coordinate space. This decoder is used only during the evaluation phase and does not participate in any training process, thus avoiding information leakage. To directly quantify the model’s contextual reasoning and semantic completion capabilities in high-occlusion scenarios, this paper uses the missing keypoint recovery success rate as an evaluation metric. During the GAT inference process, keypoints with a confidence score below 0.3 are restored to their spatial coordinates through an implicit feature decoding mechanism. The node feature vectors output by the GAT are transformed back to the spatial coordinate domain using a post-hoc fitted linear decoder, and the Euclidean distance to the original detection value is calculated. If the pixel error is ≤ 5 pixels, the restoration is considered successful. The proportion of all missing keypoints that meet this condition is counted and used as an indicator of context completion capability.

Figure 7 compares the success rates of missing keypoint recovery under high occlusion conditions using the proposed method, ResNet3D, I3D, and ST-GCN. This covers seven joint types: head, shoulder, elbow, wrist, hip, knee, and ankle. The proposed method

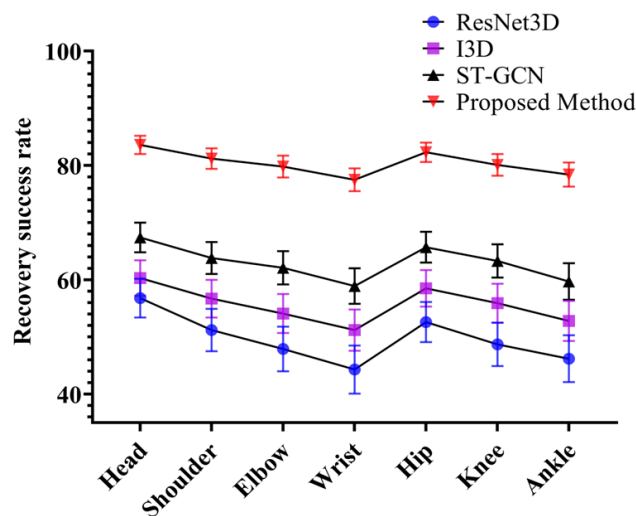


FIGURE 7. Success rate of missing keypoint recovery

significantly outperforms all joint types, achieving $83.6 \pm 1.6\%$ success rate for the head and $81.2 \pm 1.8\%$ for the shoulder. Even for the more challenging wrists and ankles, they achieve success rates of $77.5 \pm 2.0\%$ and $78.4 \pm 2.1\%$, respectively. The final average success rate is $80.4 \pm 1.9\%$, significantly higher than ResNet3D’s $49.7 \pm 3.8\%$, I3D’s $55.6 \pm 3.4\%$, and ST-GCN’s $63.0 \pm 2.9\%$. Although ST-GCN has the ability to model graph structures, it is limited by fixed adjacency relationships in dynamic interaction scenarios, and its recovery rate is still 17.4% lower than that of the method proposed in this paper, indicating that the traditional skeleton model has limited completion capabilities in complex arrangements.

When the confidence of keypoints is lower than 0.3, the proposed method does not simply rely on temporal extrapolation or geometric constraints. Instead, it adaptively weighs the state information of neighboring dancers through GAT and uses the group collaboration model to reversely infer the spatial position of the occluded individual. In the “Lift” action, when a dancer’s arm is completely obscured, the model can reconstruct the coordinates of her elbow and wrist based on the direction and relative displacement of her partner’s arm, combined with the attention weight distribution mechanism; at the same time, the global motion consistency prior provided by the C3D branch further constrains the decoding space, enabling the model to achieve a high keypoint recovery rate without an explicit reconstruction head.

3.7. Inference time and resource usage. In order to comprehensively evaluate the real-time performance and computational efficiency of the model in practical applications, this paper quantitatively analyzes the end-to-end inference time and GPU memory usage of each method. The experiment is conducted in a stable environment without interference from other processes, and the built-in timer of PyTorch is used to accurately measure the complete delay of a single video clip from input forward propagation to the final classification output. At the same time, the GPU memory consumed by the model during inference is recorded. All models are run on the same hardware platform and use a consistent input data format and preprocessing process to ensure fairness in performance comparison.

Figure 8 systematically compares the average inference time and resource usage of the proposed method, ResNet3D, I3D, and ST-GCN on the same hardware platform. ResNet3D is the fastest with the lowest latency of 68.3 ms. Its structure relies only on three-dimensional convolution operations, with a short and highly parallelized computation path and the lowest GPU memory usage (1,024 MB). However, this method does not model the structural relationships of the human body, and its performance degrades significantly under complex occlusions. I3D has an average inference time of 95.7 ms, the

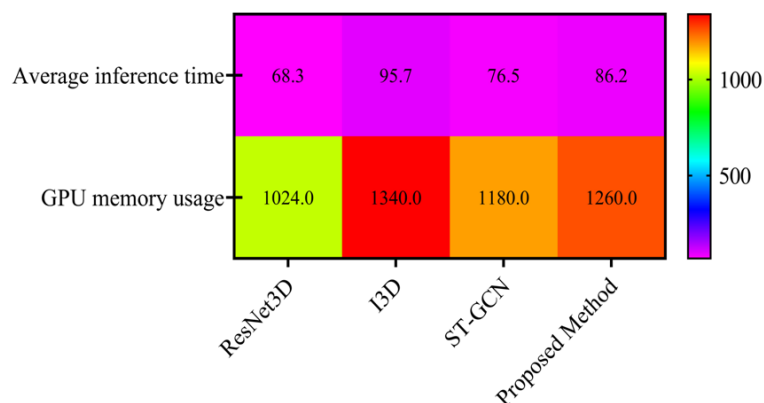


FIGURE 8. Analysis of inference time (ms) and resource usage (MB)

longest among all methods, and the GPU memory usage is also the highest (1,340 MB). ST-GCN extracts skeleton features through graph convolution, taking 76.5 ms and outperforming I3D. The average time consumption of this method is 86.2 ms, which is higher than ResNet3D and ST-GCN, but still lower than I3D. It also has complete end-to-end reasoning capabilities and is suitable for online analysis tasks in actual dance scenes.

While maintaining high discriminative performance, this method achieves a balance between efficiency and expressiveness through the collaborative optimization of the computing architecture. The C3D branch uses a compact three-dimensional convolutional structure to complete the parallel extraction of spatiotemporal features within a limited depth, avoiding redundant calculations; although the GAT branch applies a graph attention mechanism, its modeling object is dancer-level nodes, and the graph scale is much smaller than the pixel or joint granularity, which significantly reduces the complexity of adjacency matrix operations and message passing. The multi-head attention mechanism implements neighbor aggregation in the form of a matrix, without the need for serialization processing, and gets rid of the serial dependency bottleneck inherent in time series models such as LSTM. Both dual-stream paths can be executed in parallel in a single forward reasoning, and the fusion process only involves feature splicing and fully connected transformations, without iterative operations, achieving low-latency end-to-end reasoning while ensuring context-awareness. This design takes account of both the rich representation of action semantics and real-time requirements, providing a solution with engineering potential for continuous action recognition in high-density interaction scenarios.

3.8. Ablation experiment performance gain. To systematically evaluate the contribution of each core component to the overall performance, this paper designs a step-by-step ablation experiment. Under the same training and test conditions, starting from the full model, individual modules are removed one by one, and the remaining configurations are fixed. The specific variants are as follows. 1) The C3D branch input is disabled, and only the GAT branch output is used for classification. 2) The graph attention network is disabled, and action recognition is performed only through the C3D branch. 3) The feature splicing operation is canceled; the two stream branches independently output prediction results, and the probability average is taken as the final decision. 4) The attention mechanism is disabled in the GAT module; the aggregation weights of neighbor nodes are set to be uniformly distributed; the rest of the structure remains unchanged. All variants use the same network depth, learning rate, training cycle, and data augmentation strategy as the full model to ensure fair comparison. The accuracy and Macro-F1 score of each variant are evaluated on a unified test set. The independent contribution of each component is quantified by the degree of performance degradation, verifying the effectiveness of the multimodal fusion and dynamic context modeling in the method.

Table 3 compares the impact of different modules of the proposed method on recognition performance. The full model, with all components enabled, achieves the highest accuracy

TABLE 3. Ablation experiment table

Model variant	C3D branch	GAT context reasoning	Feature fusion	Attention mechanism	Accuracy (%)	Macro-F1 (%)	p-value (vs. full model)
Full model	✓	✓	✓	✓	87.3 ± 1.4	86.7 ± 1.3	–
	–	✓	✓	✓	79.4 ± 1.6	78.2 ± 1.5	$p < 0.001$
	✓	–	✓	✓	80.9 ± 1.5	79.6 ± 1.6	$p < 0.001$
	✓	✓	–	✓	83.1 ± 1.4	82.0 ± 1.4	$p < 0.005$
	✓	✓	✓	–	84.3 ± 1.3	83.5 ± 1.4	$p < 0.01$

of $87.3 \pm 1.4\%$ and Macro-F1 of $86.7 \pm 1.3\%$, significantly outperforming all variants. When the C3D branch is removed, the accuracy drops to $79.4 \pm 1.6\%$, and Macro-F1 drops to $78.2 \pm 1.5\%$, respectively. These differences are highly significant compared to the full model ($p < 0.001$). Disabling GAT context reasoning also significantly reduces performance to $80.9 \pm 1.5\%$ and $79.6 \pm 1.6\%$ ($p < 0.001$). While disabling feature fusion or the attention mechanism has relatively minor impacts, they still result in accuracy decreases of 4.2% and 3.0%, respectively. These statistically significant decreases validate the effectiveness of multimodal deep fusion and the adaptive weighting mechanism.

The spatiotemporal convolutional features extracted by the C3D branch can stably capture the overall limb movement trend, retaining rhythm and direction information even in the presence of cluttered backgrounds or partial occlusions; the GAT branch explicitly models the collaboration and interaction patterns between dancers through dynamic graph structures, allowing the actions of occluded individuals to be inferred from the states of neighboring dancers; the feature splicing and fusion strategy achieves the complementary advantages of visual and structural modalities, avoiding the discrimination bias caused by independent decision-making; the attention mechanism gives the model the ability of “selective attention”. In key actions such as “Lift” or “Crossing Position Swap”, the system can automatically focus on the most behaviorally relevant collaborators, rather than treating all neighbors equally, thereby improving the accuracy of contextual reasoning.

To verify the rationality of the GAT architecture design, additional ablation experiments were conducted on the number of layers and attention heads, as shown in Table 4. Using a single GAT layer achieves an accuracy of $84.6 \pm 1.5\%$, which is 2.7% lower than the full configuration, indicating that a single layer lacks sufficient capacity to model complex multi-dancer interaction patterns. Keeping two GAT layers but reducing attention heads from 8 to 4 results in an accuracy drop to $85.8 \pm 1.4\%$, suggesting that fewer heads limit the model’s ability to capture diverse relational features from different subspaces. Increasing to three GAT layers yields an accuracy of $86.9 \pm 1.6\%$, slightly lower than the two-layer configuration, potentially due to over-smoothing where node features become indistinguishable after excessive message passing. The experimental results confirm that the two-layer GAT with eight attention heads achieves the optimal balance between modeling capacity and generalization performance, providing empirical justification for the architectural choices in this paper.

TABLE 4. Ablation study on GAT configuration

GAT layers	Attention heads	Accuracy (%)	Macro-F1 (%)
1 layer	8 heads	84.6 ± 1.5	83.9 ± 1.4
2 layers	4 heads	85.8 ± 1.4	85.1 ± 1.5
2 layers	8 heads	87.3 ± 1.4	86.7 ± 1.3
3 layers	8 heads	86.9 ± 1.6	86.2 ± 1.5

4. Conclusion. This paper proposes a two-stream action recognition framework that integrates C3D and graph attention networks to address the challenges of high occlusion and contextual discontinuity in complex dance choreography. The C3D branch extracts global spatiotemporal motion features while constructing a dynamic human relationship graph based on confidence-weighted keypoints. GAT is used to adaptively aggregate neighborhood information to enable action inference for occluded individuals. Feature concatenation and attention pooling are then used to achieve multimodal fusion classification. Experiments demonstrate that this method achieves an average accuracy of $87.3 \pm 1.4\%$ for five typical dance action categories. Performance remains stable in high occlusion

scenarios, with a drop of only 7.8%, effectively improving the robustness and accuracy of action recognition in complex choreography. Nevertheless, several limitations point towards promising future directions. Firstly, the model's performance is contingent upon the quality of the OpenPose detections; significant errors in keypoint localization under extreme visual degradation could propagate into the graph structure. Future work will explore integrating pose estimation within an end-to-end trainable framework to enable joint optimization. Secondly, the current dataset primarily consists of professional performance and rehearsal scenes, covering a limited range of dance styles including modern dance, folk group dance, and ballet. The generalization capability of the proposed method to other dance genres with fundamentally different movement vocabularies – such as street dance, ballroom dancing, or traditional cultural dances with distinct gesture systems – requires further validation. Expanding the dataset to include a wider diversity of dance styles, recording conditions, and real-world application scenarios is a planned next step to comprehensively assess the model's robustness and adaptability. Finally, while the compact C3D backbone ensures efficiency, replacing it with more powerful architectures like I3D or VideoSwin could potentially unlock higher performance ceilings, which warrants comprehensive investigation.

REFERENCES

- [1] Y. Pang and Y. Niu, Dance video motion recognition based on computer vision and image processing, *Applied Artificial Intelligence*, vol.37, no.1, 2226962, 2023.
- [2] Z. Yue, Investigating the application of computer VR technology and AI in the examination of dance movement patterns, *Journal of Ambient Intelligence and Humanized Computing*, vol.16, no.4, pp.629-640, 2025.
- [3] Y. Zhong, X. Fu, Z. Liang et al., The application of artificial intelligence technology in the field of dance, *Applied System Innovation*, vol.8, no.5, 127, 2025.
- [4] N. Li and S. Boers, Human motion recognition in dance video images based on attitude estimation, *Wireless Communications and Mobile Computing*, vol.2023, no.1, 4687465, 2023.
- [5] T. Tang and H. Min, Research on sports dance movement detection based on pose recognition, *Mathematical Problems in Engineering*, vol.2022, no.1, 4755127, 2022.
- [6] H. Kao, Multi-person dance tiered posture recognition with cross progressive multi-resolution representation integration, *PLoS ONE*, vol.19, no.6, e0300837, 2024.
- [7] J. Feng, R. Zhang, Y. Wang et al., PoseDucer: Implicit relation inducement for invisible keypoint reconstruction in real-world occluded scenes, *Neurocomputing*, vol.640, 130328, 2025.
- [8] X. Chen, J. Zhang, K. Wang et al., Multi-person 3D pose estimation with occlusion reasoning, *IEEE Transactions on Multimedia*, vol.26, pp.878-889, 2023.
- [9] P. Patel-Grosz, P. G. Grosz, T. Kelkar et al., Steps towards a semantics of dance, *Journal of Semantics*, vol.39, no.4, pp.693-748, 2022.
- [10] G. Orgs, S. Vicary, M. Sperling et al., Movement synchrony among dance performers predicts brain synchrony among dance spectators, *Scientific Reports*, vol.14, no.1, 22079, 2024.
- [11] Y. Qin, S. Liu and C. Yuan, Enhanced spatiotemporal skeleton modeling: Integrating part-joint attention with dynamic graph convolution, *Scientific Reports*, vol.15, no.1, 34781, 2025.
- [12] L. Zhang, Improved graph convolutional neural network for dance tracking and pose estimation, *Computational Intelligence and Neuroscience*, vol.2022, no.1, 7133491, 2022.
- [13] J. Sun and L. Song, Dance movement pose estimation in complex scenes based on improved high-resolution networks, *Journal of Network Intelligence*, vol.10, no.1, pp.410-423, 2025.
- [14] D. Sarma, V. Kavyasree and M. K. Bhuyan, Two-stream fusion model using 3D-CNN and 2D-CNN via video-frames and optical flow motion templates for hand gesture recognition, *Innovations in Systems and Software Engineering*, vol.21, no.1, pp.39-52, 2025.
- [15] Q. Xu, W. Zheng, Y. Song et al., Scene image and human skeleton-based dual-stream human action recognition, *Pattern Recognition Letters*, vol.148, pp.136-145, 2021.
- [16] L. Tian, Z. Yang, J. Yang et al., Spatial-temporal attention augmented graph convolution method based on human posture response, *Sensors and Actuators A: Physical*, vol.395, 117017, 2025.
- [17] J. Qu, A dance movement quality evaluation model using transformer encoder and convolutional neural network, *Scientific Reports*, vol.14, no.1, 32058, 2024.

- [18] L. Wu, X. Lang, Y. Xiang et al., Multi-perspective representation to part-based graph for group activity recognition, *Sensors*, vol.22, no.15, 5521, 2022.
- [19] L. Li and T. Yang, Reconstruction of physical dance teaching content and movement recognition based on a machine learning model, *3C TIC. Cuadernos de desarrollo aplicados a las TIC*, vol.12, no.1, pp.267-285, 2023.
- [20] M. Li, Ethnic dance movement recognition based on motion capture sensor and machine learning, *International Journal of Information and Communication Technology*, vol.25, no.8, pp.81-96, 2024.
- [21] Q. Lei, H. Zhang and J. Du, Temporal attention learning for action quality assessment in sports video, *Signal, Image and Video Processing*, vol.15, no.7, pp.1575-1583, 2021.
- [22] H. Xu, X. Jin, Q. Wang et al., Exploiting attention-consistency loss for spatial-temporal stream action recognition, *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol.18, no.2s, pp.1-15, 2022.
- [23] L. Zhu, Z. Wang, S. Liu et al., Multi-person pose regression with distribution-aware single-stage models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.46, no.8, pp.5384-5397, 2024.
- [24] L. Jin, X. Wang, X. Nie et al., Rethinking the person localization for single-stage multi-person pose estimation, *IEEE Transactions on Multimedia*, vol.26, pp.1436-1447, 2023.
- [25] Z. Jin, Y. Wang, Q. Wang et al., SSRL: Self-supervised spatial-temporal representation learning for 3D action recognition, *IEEE Transactions on Circuits and Systems for Video Technology*, vol.34, no.1, pp.274-285, 2023.
- [26] F. T. Peng and H. Zhang, Research on action recognition method of dance video image based on human-computer interaction, *Scientific Programming*, vol.2021, no.1, 8763133, 2021.
- [27] M. L. Gau, H. Y. Ting, J. T. W. Ting et al., Sarawak traditional dance motion analysis and comparison using Microsoft Kinect V2, *Green Intelligent Systems and Applications*, vol.2, no.1, pp.42-52, 2022.
- [28] X. Ju, The application of deep learning in dance movement design, *International Journal of Computational Intelligence Systems*, vol.18, no.1, pp.1-23, 2025.
- [29] M. Wang, X. Li, S. Chen et al., Learning representations by contrastive spatio-temporal clustering for skeleton-based action recognition, *IEEE Transactions on Multimedia*, vol.26, pp.3207-3220, 2023.
- [30] M. Rahevar, A. Ganatra, T. Saba et al., Spatial-temporal dynamic graph attention network for skeleton-based action recognition, *IEEE Access*, vol.11, pp.21546-21553, 2023.
- [31] A. Bera, M. Nasipuri, O. Krejcar et al., Fine-grained sports, yoga, and dance postures recognition: A benchmark analysis, *IEEE Transactions on Instrumentation and Measurement*, vol.72, pp.1-13, 2023.
- [32] T. D. Itoh, T. Kubo and K. Ikeda, Multi-level attention pooling for graph neural networks: Unifying graph representations with multiple localities, *Neural Networks*, vol.145, pp.356-373, 2022.
- [33] W. Ali, S. Vascon, T. Stadelmann et al., Hierarchical glocal attention pooling for graph classification, *Pattern Recognition Letters*, vol.186, pp.71-77, 2024.
- [34] Q. Lv, S. Zhang and Y. Wang, Deep learning model of image classification using machine learning, *Advances in Multimedia*, vol.2022, no.1, 3351256, 2022.
- [35] H. Alaeddine and M. Jihene, Deep residual network in network, *Computational Intelligence and Neuroscience*, vol.2021, no.1, 6659083, 2021.

Author Biography



Jingya Chen earned her Ph.D. degree in Education from Sehan University, Korea, in 2023. She currently serves as a Dance Instructor at the School of Education and Music, Hainan Vocational University of Science and Technology, China. Her research interests include dance education, early childhood dance education, and information-based teaching. She has been awarded the International Certified Senior Professional Model Trainer Certificate by the American ACIC International Professional Skills Certification Examination. She has led students to win awards in provincial and ministerial-level competitions on multiple occasions. She has published eight papers, one monograph, one school-enterprise cooperation textbook at the university level, and is currently leading one provincial-level research project and one key university-level project.



Qiong Yang received her master's degree from East China Normal University, China, in 2012. She currently engages in College of Information Technology, Qiongtai Normal University, China, where she serves as the Academic Discipline Leader and the Director of the Educational Technology Department. Her research interests include information technology and multimedia applications, ICT in basic education, the cultivation of information technology application skills for normal university students, and information technology curriculum and instruction in primary schools. She has led multiple provincial and ministerial-level research projects, completed several provincial and ministerial-level projects funded by the Natural Science Foundation, and published over 30 academic papers. She has also guided students to win numerous national micro-course awards.



Yanni Lin earned her Ph.D. degree from Sehan University, Korea, in 2024. Currently, she is an Associate Professor and a full-time Dance Instructor at Qiongtai Normal University, China, where she also serves as the Director of the Art Practice Center at the Conservatory of Music. Her research interests include dance choreography, movement composition, and information-based teaching. Her academic affiliations and honors include being a member of the Chinese Dancers Association, a member of the Chinese Ethnic Minorities Dancers Association, a council member of the Hainan Dancers Association, and a council member of the Hainan Art Education Promotion Association. She was awarded the Silver Award at the Hainan College Teacher Skills Exhibition. She has published two books and several papers, and has led one provincial-level research project. Additionally, she has participated in five provincial-level projects, one provincial-level excellent course, one university-level textbook project, and one horizontal research project.