

STME-GCN: topology refinement and motion-enhanced temporal modeling for skeleton-based action recognition

Xiaopeng Wang

Southwest Jiaotong University

Chunjun Chen

c.jchen@swjtu.edu.cn

Southwest Jiaotong University

Yutao Xia

Southwest Jiaotong University

Research Article

Keywords: keleton-based action recognition, graph convolutional networks, motion enhancement, dynamic topology

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10937561/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

STME-GCN: topology refinement and motion-enhanced temporal modeling for skeleton-based action recognition

Xiaopeng Wang^{1,2}, Chunjun Chen^{1,3*}, Yutao Xia^{1,3}

¹School of Mechanical Engineering, Southwest Jiaotong University, Chengdu, 610031, China.

²Tangshan Research Institute, Southwest Jiaotong University, Tangshan, 063000, China.

³Technology and Equipment of Rail Transit Operation and Maintenance Key Laboratory of Sichuan Province, Chengdu, 610031, China.

*Corresponding author(s). E-mail(s): cjchen@swjtu.edu.cn;
Contributing authors: xpwang@my.swjtu.edu.cn; xytpxl@swjtu.edu.cn;

Abstract

In recent years, graph convolutional networks (GCNs) have been widely applied to skeleton-based action recognition. However, existing channel-wise dynamic topology methods typically compress node features along the temporal dimension before topology generation, which limits the use of feature variations, motion intensity, and phase evolution for estimating inter-joint relations. Meanwhile, conventional multi-scale temporal convolutions mainly model temporal dependencies by aggregating features within local temporal windows, without explicitly characterizing inter-frame variations, thereby limiting the representation of fine-grained motion dynamics. To address these issues, we propose STME-GCN, a skeleton-based action recognition network featuring topology feature optimization and temporal motion enhancement. Specifically, the Topology Feature Optimization Module integrates multi-attribute temporal descriptors, including global state, local motion, and phase evolution, to enrich the node information used for generating channel-specific topologies. The Temporal Motion Enhancement Module introduces a pose–motion collaborative branch alongside the multi-scale temporal pathway to jointly extract pose features and explicit motion features, thereby enhancing the representation of local temporal dynamics. Experimental results demonstrate that, with relatively few parameters, STME-GCN achieves competitive recognition accuracy on NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA, providing a favorable balance between model efficiency and recognition performance.

Keywords: skeleton-based action recognition; graph convolutional networks; motion enhancement; dynamic topology

1 Introduction

Action recognition aims to determine action categories from videos or motion sequences and has important applications in video surveillance,

human–computer interaction, medical rehabilitation, and sports analysis. RGB-based methods can fully exploit appearance, texture, and scene information, but their recognition performance is susceptible to changes in illumination, background, viewpoint, and occlusion. In contrast,

human skeleton data describe body motion using joint coordinates and their connectivity, providing low-dimensional representations with clear semantic structures [1]. Early skeleton-based methods mainly employed recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to process skeleton sequences, making it difficult to explicitly model the inherent non-Euclidean structure among human joints [1, 2]. ST-GCN first represented human skeletons as spatio-temporal graphs and jointly extracted action features through spatial graph convolution and temporal convolution [2]. Building on this framework, subsequent studies have continuously improved skeleton representations through adaptive topology learning [3, 4], channel-wise relation modeling [5], and multi-scale spatio-temporal feature extraction [6]. In particular, CTR-GCN generates channel-specific relation refinement terms according to pairwise node feature differences, allowing different feature channels to adopt distinct information propagation patterns [5]. Nevertheless, effectively exploiting the temporal evolution information contained in action sequences to further improve inter-joint relation modeling and temporal representations remains an important problem in skeleton-based action recognition.

Despite the favorable performance achieved by existing methods, limitations remain in topology generation and local motion modeling. Channel-wise dynamic topology methods commonly aggregate node features along the temporal dimension and then calculate inter-joint correlations from the compressed representations [5]. Although this strategy provides efficient topology inference, temporal averaging mainly describes the overall state of an action sequence and cannot sufficiently preserve feature fluctuations, local motion variations, or phase evolution. Some methods retain fine-grained dynamics by explicitly constructing time-varying inter-joint relationships [7], which introduces topology modeling along the temporal dimension. Therefore, the first problem addressed in this work is how to enable channel-wise topology modeling to exploit richer action-evolution information without explicitly constructing frame-specific inter-joint relations.

In terms of temporal modeling, multi-scale temporal modeling can extract action features over different temporal ranges [6, 8], whereas local motion variations between adjacent frames are

generally learned only implicitly by the network. Explicit motion information is usually introduced through additional input streams, such as joint motion and bone motion [9], whose corresponding feature extraction networks and classification branches further increase model complexity. Accordingly, the second problem considered in this work is how to make full use of action evolution information without constructing frame-wise dynamic topologies or introducing an independent motion stream.

To address these problems, we propose STME-GCN, a skeleton-based action recognition network based on topology feature optimization and temporal motion enhancement. The Topology Feature Optimization (TFO) Module replaces a single temporal-mean descriptor with multi-attribute temporal descriptors, thereby providing more informative temporal cues for inter-joint relation estimation. The Temporal Motion Enhancement (TME) Module strengthens fine-grained dynamic features by explicitly modeling motion variations. The main contributions of this work are summarized as follows:

1. We propose a channel-wise Topology Feature Optimization Module that encodes the global state, local motion responses, and phase evolution information of joint sequences into node-level temporal representations. Input-dependent channel-wise topology refinement terms are then generated from the representation differences between nodes, mitigating the loss of temporal-evolution information caused by reliance on a single temporal-mean descriptor.
2. We propose a Temporal Motion Enhancement Module that constructs a pose-motion collaborative pathway in addition to multi-scale temporal convolution. First-order inter-frame differences of shallow-layer features are employed to explicitly emphasize local motion variations and are fused with the base temporal features in a learnable residual manner, without introducing an independent motion input stream or an additional classification branch.
3. We validate the effectiveness of the proposed method on the NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA datasets. Experimental results demonstrate that temporal-evolution-driven topology optimization and

shallow-layer motion enhancement are complementary. The complete model achieves competitive recognition performance while maintaining a relatively small number of parameters and low computational cost.

2 Related Work

2.1 Dynamic Topology Modeling

ST-GCN constructs a fixed adjacency matrix according to the natural connections of the human skeleton and aggregates features from neighboring joints through graph convolution [2]. Since a fixed topology has limited flexibility in modeling non-physical joint dependencies across different actions, 2s-AGCN further introduces learnable adjacency matrices and sample-dependent relations, enabling information propagation among joints to adapt to the input features [3]. Subsequent studies have improved the representational capacity of graph topology through graph-structure decomposition [6], hierarchical organization [10], and channel-wise differentiation [5]. CTR-GCN generates channel-specific inter-joint relations according to pairwise node feature differences, allowing different feature channels to learn distinct information propagation patterns [5]. BlockGCN and DS-GCN further extend the representation of inter-joint relations through decoupled graph structures and reorganized topology units, respectively [11, 12]. More recent studies have further explored adaptive and dynamic topology modeling through multi-dimensional graph construction [13], adaptive large-kernel graph modeling [8], dynamic topology with multi-scale temporal modeling [14], and topology-enhanced spatio-temporal graph convolution [15].

2.2 Temporal Modeling and Motion Representation

Temporal convolution is widely used for temporal feature extraction in GCN-based skeleton action recognition. Early methods employed temporal convolution kernels with fixed receptive fields to aggregate features from neighboring frames [2]. Subsequently, multi-scale temporal

modeling has been introduced to enlarge the temporal receptive field and capture action information over different temporal ranges [6, 8]. Recent studies have further explored Transformer-based and state-space architectures to model long-range spatio-temporal dependencies [16, 17]. However, these temporal modeling approaches generally extract motion dynamics from continuous feature sequences, without explicitly introducing local inter-frame differences into temporal feature extraction.

To explicitly incorporate action variations, existing studies construct inter-frame joint or bone differences as independent motion modalities and employ multi-stream networks for feature extraction and decision-level fusion [9]. Although these approaches provide complementary motion cues, they generally require additional input branches and classifiers.

3 Proposed Method

STME-GCN consists of a Topology Feature Optimization Module and a Temporal Motion Enhancement Module. The former provides more informative node representations for channel-specific topology generation through multi-attribute temporal descriptors, whereas the latter enhances shallow-layer local dynamic features through a pose-motion collaborative pathway. The following subsections introduce the two modules and describe how they are deployed within the overall network.

3.1 Overall Architecture

The proposed STME-GCN integrates the Topology Feature Optimization Module and the Temporal Motion Enhancement Module into a spatio-temporal graph convolutional backbone. The overall architecture is illustrated in Fig. 1.

The input skeleton sequence is denoted as $X_0 \in \mathbb{R}^{B \times C \times T \times N \times M}$, where M denotes the number of persons in each sequence. After data normalization and dimension rearrangement, the person dimension is merged into the batch dimension before the features are fed into the subsequent network. The backbone of STME-GCN consists of ten basic units and is divided into three stages according to the number of output channels. The

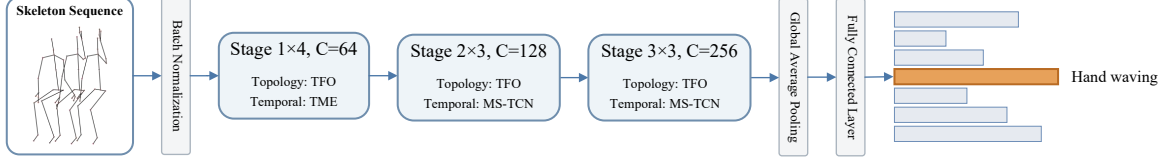


Fig. 1 Overall architecture of STME-GCN, with TFO used for topology modeling in all blocks and TME applied only to the four shallow blocks

first four units have 64 output channels, the middle three units have 128 channels, and the last three units have 256 channels. Each basic unit sequentially performs spatial graph feature aggregation and temporal feature extraction, followed by a residual connection. The Topology Feature Optimization Module is used in all ten basic units for channel-wise topology refinement at different network depths. In contrast, the Temporal Motion Enhancement Module is deployed only in the first four shallow units, where local motion variations are still relatively explicit. The subsequent layers continue temporal feature extraction using the base multi-scale temporal convolution, thereby avoiding excessive enhancement of local inter-frame variations in deep semantic features. After multi-layer spatio-temporal feature extraction, global average pooling is applied to obtain action-level representations, which are then fed into a fully connected layer for final category prediction. In this way, the Topology Feature Optimization Module improves spatial relation modeling, while the Temporal Motion Enhancement Module supplements shallow-layer local dynamic information. Together, they constitute the spatio-temporal feature extraction framework of STME-GCN.

3.2 Topology Feature Optimization (TFO) Module

Existing channel-specific dynamic topology methods typically aggregate node features along the temporal dimension before computing inter-joint relations from the compressed representations [5]. Although efficient, a single temporal mean mainly reflects the overall sequence state and may suppress local motion variations and phase-wise evolution. Temporal frame-wise topology methods explicitly construct inter-joint relations at different temporal positions to preserve fine-grained dynamic information [7], but require inter-joint

relations to be modeled across temporal positions, increasing the computational cost of topology generation. To address this trade-off, TFO retains the channel-specific topology computation paradigm while replacing the single temporal-mean representation with multi-attribute temporal descriptors that encode global state, local motion, and phase evolution. These descriptors provide richer temporal cues for inter-joint relation generation without explicitly constructing frame-wise pairwise relations. The structure of TFO is shown in Fig. 2.

Given an input feature $X \in \mathbb{R}^{B \times C \times T \times N}$, where B , C , T , and N denote the batch size, number of channels, number of frames, and number of joints, respectively, three independent 1×1 convolutions are applied to obtain

$$(P_k, Q_k, H_k) = \Phi_k(X), \quad (1)$$

where $\Phi_k(\cdot)$ denotes three independent channel projections for the k -th subgraph. Here, $P_k, Q_k \in \mathbb{R}^{B \times C_r \times T \times N}$ are used for pairwise relation modeling, while $H_k \in \mathbb{R}^{B \times C_{out} \times T \times N}$ is used for subsequent feature aggregation. Subsequently, for each relation projection feature $F \in \{P_k, Q_k\}$, three types of temporal information are extracted along the temporal dimension, namely the global state, local motion, and phase evolution. Their specific definitions are given as follows.

Ω_s denotes the set of frames corresponding to the s -th temporal phase, where $s \in \{1, 2, 3\}$. The phase differences are calculated between the early and middle phases, the middle and late phases, and the early and late phases, respectively. The global-state descriptor characterizes the overall feature state and sequence-level variation, the local-motion descriptor captures the intensity of short-term inter-frame changes, and the phase-evolution descriptor characterizes state transitions across different temporal phases of an action. After concatenating the three types of descriptors

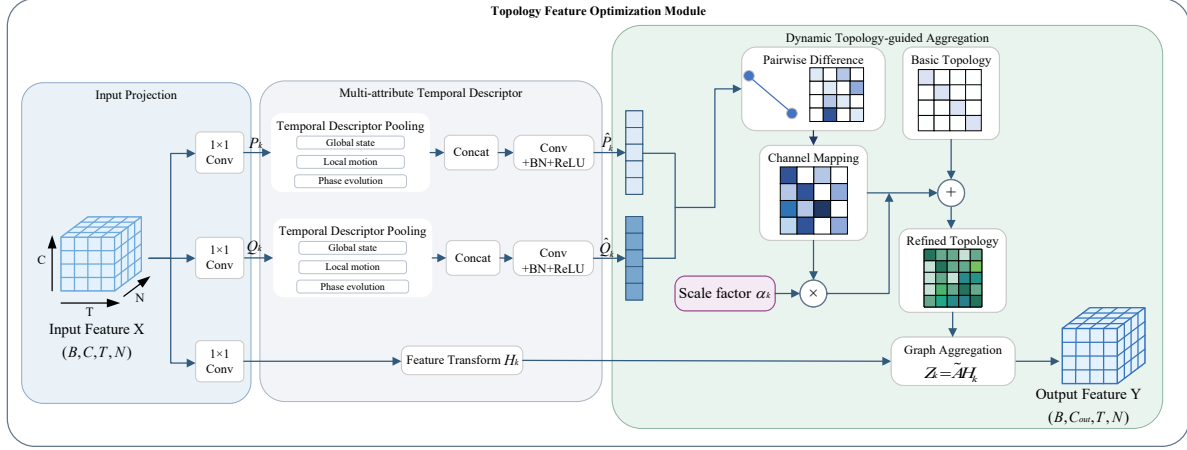


Fig. 2 Architecture of the Topology Feature Optimization (TFO) module, where multi-attribute temporal descriptors replace the single temporal-mean representation for channel-wise topology refinement

Table 1 Definitions and roles of the multi-attribute temporal descriptors in TFO

Temporal Information	Mathematical Definition	Function
Global State	$\mu(F) = \frac{1}{T} \sum_{t=1}^T F_t,$	Describes the average feature state and degree of fluctuation over the entire sequence
	$\sigma(F) = \sqrt{\frac{1}{T} \sum_{t=1}^T (F_t - \mu(F))^2 + \epsilon}$	
Local Motion	$\Delta F_t = F_t - F_{t-1}, \quad t = 2, \dots, T,$	Captures the average intensity and salient variations of short-term motion
	$M_{\text{avg}}(F) = \frac{1}{T-1} \sum_{t=2}^T \Delta F_t ,$	
	$M_{\text{max}}(F) = \max_{2 \leq t \leq T} \Delta F_t $	
Phase Evolution	$p_s(F) = \frac{1}{ \Omega_s } \sum_{t \in \Omega_s} F_t, \quad s \in \{1, 2, 3\},$	Characterizes feature-state differences and evolution across different action phases
	$d_{21} = p_2(F) - p_1(F),$	
	$d_{32} = p_3(F) - p_2(F),$	
	$d_{31} = p_3(F) - p_1(F)$	

along the channel dimension, a 1×1 convolution followed by batch normalization and ReLU is applied to remap them into the C_r -dimensional relation space, yielding the node-level temporal representations $\hat{P}_k$ and $\hat{Q}_k$.

After obtaining the node-level temporal representations, the relations between source and target nodes are computed through pairwise broadcast subtraction. For nodes u and v , the relation

feature is defined as

$$R_{k,b,c,u,v} = \tanh(\hat{P}_{k,b,c,u} - \hat{Q}_{k,b,c,v}), \quad (2)$$

where $R_k \in \mathbb{R}^{B \times C_r \times N \times N}$. This operation generates input-dependent pairwise node relations according to the representation differences between nodes in terms of global state, local motion, and phase evolution. Subsequently, a channel mapping function $g_k(\cdot)$ projects the relation features into the output channel space and fuses them with the k -th base sub-topology A_k :

$$\tilde{A}_k = A_k + \alpha_k g_k(R_k), \quad (3)$$

where α_k is a learnable scaling coefficient that controls the influence of the input-dependent relations on the base topology. The updated topology $\tilde{A}_k$ is then used to aggregate the node features H_k . The aggregation results from the three sub-topology branches are summed and subsequently processed by batch normalization and a residual connection to obtain the module output. In this way, multi-attribute temporal information directly participates in the generation of channel-specific topologies without explicitly constructing pairwise node relations at every temporal step, thereby providing richer temporal cues for topology generation while maintaining relatively low computational complexity.

3.3 Temporal Motion Enhancement (TME) Module

Conventional multi-scale temporal convolution aggregates consecutive-frame features using different temporal receptive fields and can effectively capture multi-scale temporal information [6, 8]. However, local variations between adjacent frames are still mainly learned implicitly by the network. Such short-term dynamics provide important discriminative cues for actions with similar poses but different motion processes. Therefore, we design a Temporal Motion Enhancement Module that introduces a pose-motion collaborative pathway alongside the multi-scale temporal pathway, explicitly incorporating inter-frame variations into shallow-layer temporal feature extraction. The module structure is illustrated in Fig. 3.

Given the output of spatial graph convolution, $Y \in \mathbb{R}^{B \times C_{in} \times T \times N}$, the base temporal pathway

consists of four parallel branches. Two temporal convolution branches use dilation rates of 1 and 2, respectively, to extract local features over different temporal ranges. A max-pooling branch supplements locally salient responses, while a 1×1 projection branch preserves the original feature information. After channel mapping, the outputs of all branches are concatenated along the channel dimension to form the base multi-scale temporal feature F_{base} .

On this basis, the pose-motion collaborative pathway explicitly extracts variations between adjacent frames. For the input feature Y , the signed first-order difference is defined as

$$\Delta Y_t = \begin{cases} 0, & t = 1, \\ Y_t - Y_{t-1}, & 2 \leq t \leq T, \end{cases} \quad (4)$$

where ΔY preserves both the direction and magnitude of feature changes between adjacent frames. Subsequently, the original pose feature Y and the motion feature ΔY are separately processed by depthwise temporal convolutions to extract local pose variations and motion responses. The two feature streams are then concatenated along the channel dimension, followed by a pointwise convolution for feature interaction and channel mapping, yielding the pose-motion fused feature F_m . Compared with relying solely on continuous pose features, this pathway directly takes inter-frame variations as input, enabling shallow network layers to exploit local dynamic information more effectively.

Finally, the pose-motion feature is fused with the base multi-scale temporal feature through a learnable residual formulation:

$$Y_{tem} = F_{base} + \alpha F_m, \quad (5)$$

where α is a learnable scaling coefficient initialized to zero. At the beginning of training, the module mainly preserves the feature propagation of the base temporal pathway. As training proceeds, the model can adaptively adjust the contribution of motion information according to the recognition objective. Since shallow features retain clearer motion semantics than deep high-level semantic features, the Temporal Motion Enhancement Module is deployed only in shallow network layers

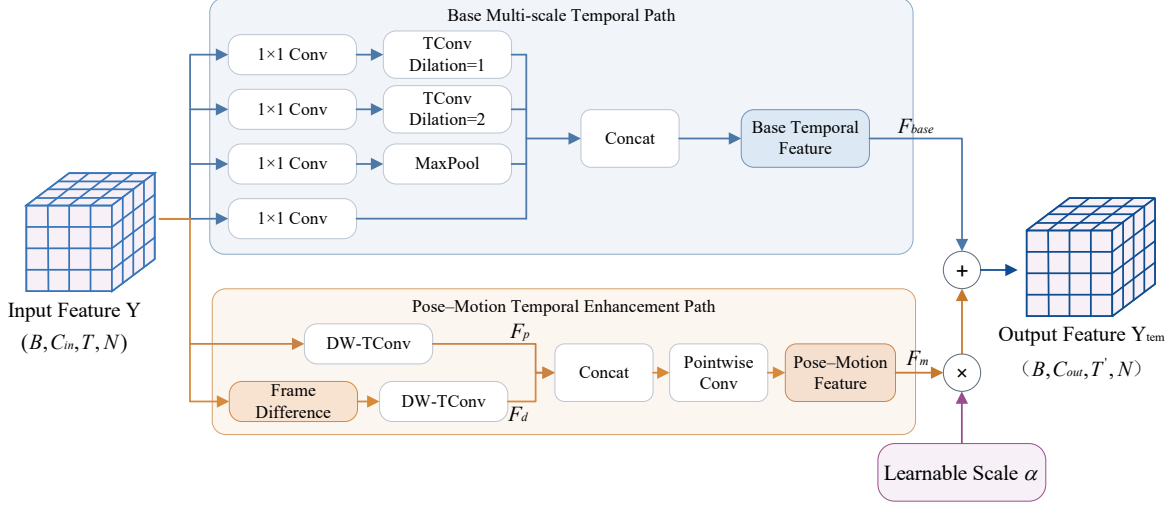


Fig. 3 Architecture of the Temporal Motion Enhancement (TME) module, which reformulates the shallow temporal block by combining the base multi-scale pathway with explicit inter-frame motion modeling

to avoid excessive amplification of local variations in deeper representations.

4 Experiments and Results

4.1 Datasets and Implementation Details

Experiments were conducted on three skeleton-based action recognition datasets: NTU RGB+D 60 [18], NTU RGB+D 120 [19], and NW-UCLA [20]. NTU RGB+D 60 contains 56,880 skeleton sequences covering 60 action classes performed by 40 subjects. Each skeleton consists of 25 joints, and the dataset is evaluated under the cross-subject (X-Sub) and cross-view (X-View) protocols. NTU RGB+D 120 further expands the dataset to 114,480 sequences, 120 action classes, and 106 subjects, and adopts the cross-subject (X-Sub) and cross-setup (X-Set) evaluation protocols. NW-UCLA contains 1,494 sequences from 10 action classes captured from three viewpoints, with each skeleton consisting of 20 joints. Following the standard cross-view protocol, samples from the first two viewpoints are used for training, while those from the third viewpoint are used for testing.

The model was implemented in PyTorch and trained using stochastic gradient descent. The momentum, weight decay, and initial learning rate were set to 0.9, 0.0004, and 0.1, respectively,

with linear warm-up applied during the first five epochs. For NTU RGB+D 60 and NTU RGB+D 120, the batch size was set to 128, and the models were trained for 120 epochs. The learning rate was reduced by a factor of 0.1 at epochs 60, 90, and 110. For NW-UCLA, the batch size was set to 16, and the models were trained for 100 epochs, with the learning rate adjusted at epochs 50, 70, and 85. Joint-stream and bone-stream models were trained separately, and their prediction scores were fused at the decision level to obtain the two-stream results. Top-1 accuracy was used as the evaluation metric in all experiments.

4.2 Comparison with Existing Methods

To evaluate the overall performance of STME-GCN, comparisons were conducted with representative methods on NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA. The top-1 accuracies of the joint stream, bone stream, and two-stream fusion are reported in Tables 2–3. Results for the competing methods were collected from the original publications, the public benchmarks provided by PYSKL [21], or the retrained baseline results reported in SkeletonGCL [26]. On NTU RGB+D 60, STME-GCN achieves 90.7% accuracy with the joint stream under the X-Sub protocol, while the two-stream model reaches 92.6%

Table 2 Comparison of Top-1 accuracy (%) with representative methods on NTU RGB+D 60 and NTU RGB+D 120

Method	NTU RGB+D 60 (X-Sub)			NTU RGB+D 60 (X-View)			NTU RGB+D 120 (X-Sub)			NTU RGB+D 120 (X-Set)		
	Joint	Bone	2s	Joint	Bone	2s	Joint	Bone	2s	Joint	Bone	2s
ST-GCN [†] [2]	87.8	88.6	90.0	95.5	95.0	96.2	82.1	83.7	85.6	84.5	85.8	87.5
AAGCN [†] [9]	89.0	89.2	90.8	95.7	95.2	96.4	82.8	84.7	86.3	84.8	86.2	88.1
2s-AGCN* [3]	88.9	89.2	91.0	94.5	94.1	95.7	84.0	85.1	87.8	85.3	86.3	89.0
CTR-GCN [†] [5]	89.6	90.0	91.5	95.6	95.4	96.6	84.0	85.9	87.5	85.9	87.4	89.2
MS-G3D [†] [6]	89.6	89.3	91.0	95.9	95.0	96.4	84.0	85.3	86.9	86.0	87.3	88.9
STGCN++ [†] [21]	89.3	90.1	91.4	95.6	95.5	96.7	83.2	85.6	87.0	85.6	87.5	89.1
H ² GCN [22]	–	–	92.2	–	–	96.1	–	–	86.9	–	–	88.9
DPGCN [23]	90.1	90.4	91.9	–	–	96.5	–	–	87.5	–	–	89.3
AT-Shift-GCN [24]	–	–	91.2	–	–	96.8	–	–	86.8	–	–	88.2
STME-GCN	90.7	90.6	92.6	96.0	95.3	96.9	85.5	85.9	88.4	87.2	87.3	90.0

Note: Joint and Bone denote the joint and bone streams, respectively; 2s denotes two-stream fusion. “–” indicates that the corresponding result was not reported. [†] denotes results obtained from the public benchmarks provided by PYSKL [21]; * denotes results reported in SkeletonGCL [26], where the corresponding models were retrained using their officially released code.

Table 3 Comparison of Top-1 accuracy (%) with representative methods on NW-UCLA

Method	NW-UCLA		
	Joint	Bone	2s
AT-Shift-GCN [24]	93.4	–	94.8
Shift-GCN [25]	92.5	–	94.2
2s-AGCN* [3]	92.0	92.2	95.0
CTR-GCN* [5]	94.6	91.8	94.2
STME-GCN	94.3	93.5	95.3

Note: * denotes retrained results reported in SkeletonGCL [26], as defined in Table 2.

and 96.9% under X-Sub and X-View, respectively. The improvement of the joint stream indicates that the proposed topology refinement and motion enhancement effectively exploit temporal evolution and local dynamics. Although the gain of the bone stream is relatively limited, joint and bone representations remain complementary, leading to improved two-stream performance. On NTU RGB+D 120, the two-stream accuracies reach 88.4% and 90.0% under X-Sub and X-Set, respectively. On NW-UCLA, STME-GCN achieves a two-stream accuracy of 95.3%. Overall, STME-GCN maintains competitive performance across datasets of different scales and evaluation protocols, with consistent advantages in the two-stream fusion.

4.3 Ablation Studies

To verify the effectiveness of the proposed modules, joint-stream ablation experiments were conducted on the NTU RGB+D 60 dataset under the X-Sub protocol, with the results reported in Table 4.

The baseline model achieves a Top-1 accuracy of 89.60%. Replacing the baseline topology modeling with TFO increases the accuracy to 89.95%, while reformulating the shallow temporal modeling with TME improves it to 90.54%. When both proposed modules are adopted, STME-GCN achieves 90.73%, outperforming the baseline by 1.13 percentage points.

In terms of model complexity, STME-GCN contains 1.61 M parameters and requires 2.00 GFLOPs. These results indicate that TFO and TME improve inter-joint relation modeling and local temporal dynamics, respectively, while maintaining relatively low computational complexity. Their joint use yields the best recognition accuracy, demonstrating the complementarity of the two designs.

Table 4 Ablation study of TFO and TME on NTU RGB+D 60 under the X-Sub protocol

Configuration	Top-1 (%)	Params. (M)	GFLOPs
Baseline	89.60	1.45	1.94
TFO variant	89.95	1.58	1.89
TME variant	90.54	1.42	1.99
STME-GCN	90.73	1.61	2.00

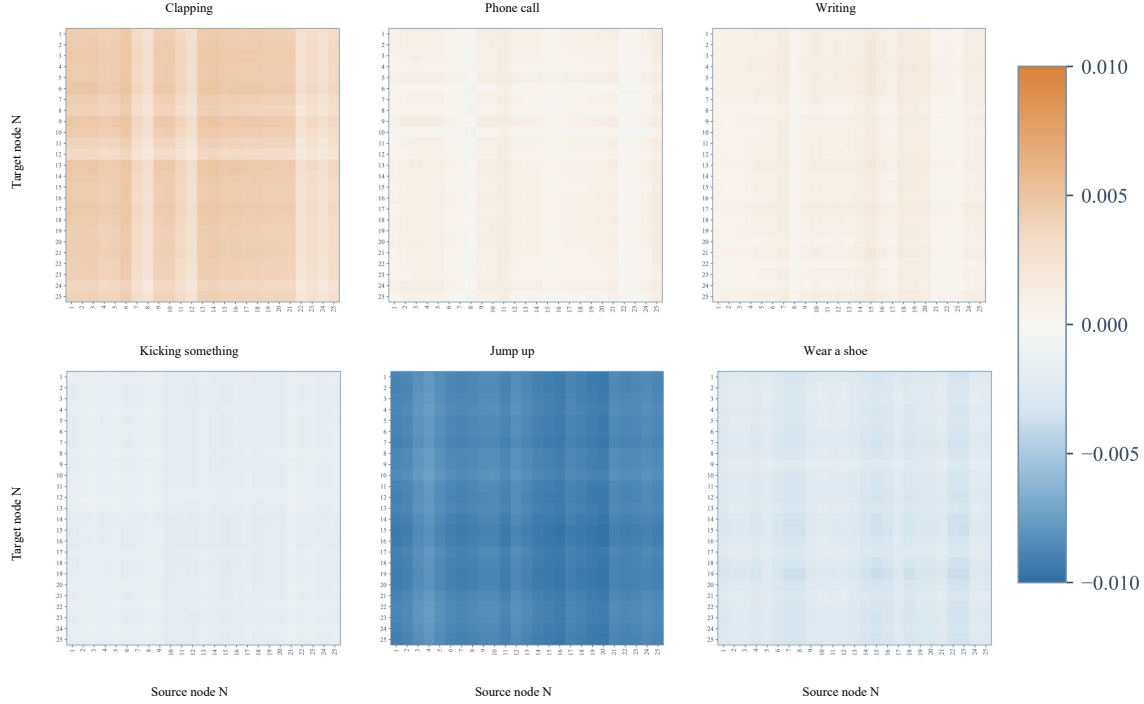


Fig. 4 Visualization of dynamic relation components learned in deep network layers for different representative actions

4.4 Visualization Analysis

To further examine the behavior of the proposed modules, we conduct visualization analyses from two complementary perspectives: topology refinement and temporal motion responses. All results are obtained using the joint-stream model on the NTU RGB+D 60 dataset.

4.4.1 Analysis of Action-Related Dynamic Topologies

To examine whether the inter-joint relations learned by the model are action-dependent, six representative action classes are selected: “Clapping,” “Phone call,” “Writing,” “Kicking something,” “Jump up,” and “Wear a shoe.” The dynamic relation components generated from the input features in the deep network layers are visualized in Fig. 4. The horizontal and vertical axes represent the source and target nodes, respectively. Positive and negative responses indicate that the model strengthens and weakens the corresponding inter-joint relations in the base topology, respectively, whereas responses close to

zero indicate only minor relation adjustments. Different actions exhibit clearly distinct response magnitudes and distributions of positive and negative values. The “clapping” action is dominated by relatively strong positive responses, indicating that the model strengthens information exchange among specific joints. The overall responses for “Phone call” and “Wear a shoe” are close to zero, with relation adjustments concentrated on only a few joint pairs. “Writing” and “Kicking something” mainly exhibit weak negative responses, whereas “Jump up” shows more pronounced overall negative modulation. These differences indicate that the Topology Feature Optimization Module does not apply a fixed relation adjustment to all actions. Instead, it adaptively modulates the strength of information propagation between different joints according to the input action features. These results further demonstrate that multi-attribute temporal information contributes to the generation of action-related topologies, allowing deep-layer inter-joint relations to vary with the action content.

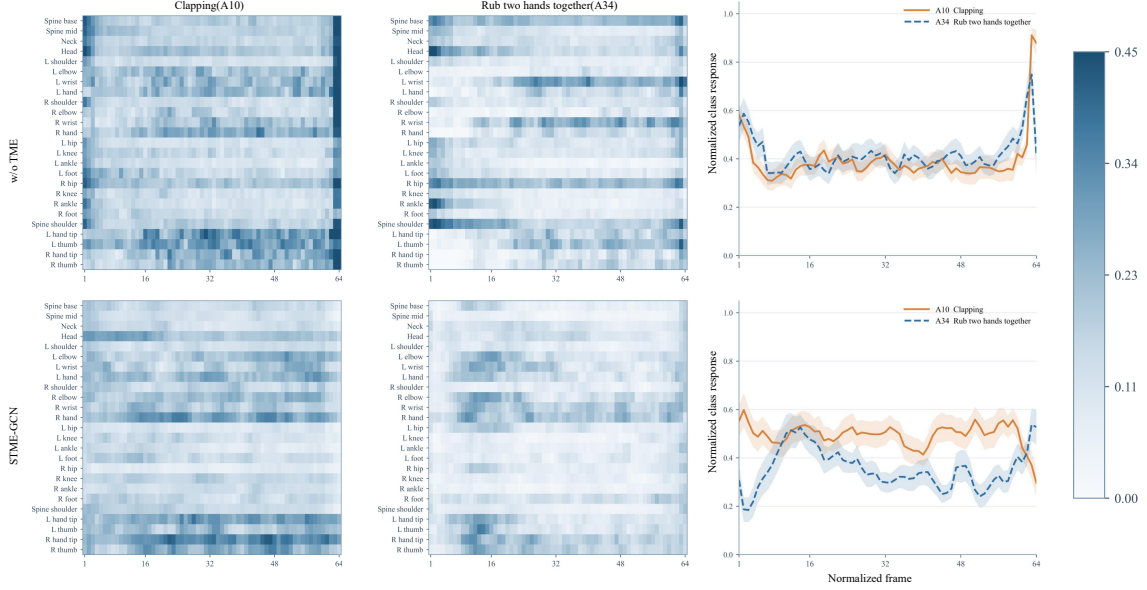


Fig. 5 Temporal response visualization for "Clapping" (A10) and "Rub two hands together" (A34) under different model configurations. The left and middle columns show the joint-time response heatmaps, while the right column presents the frame-level response curves aggregated over all joints.

4.4.2 Analysis of Temporal Motion Responses

To investigate how the Temporal Motion Enhancement Module affects local dynamic information, we select two actions, "clapping" and "rubbing hands," and compare their joint-time responses and frame-level response curves between the variant without TME, which retains TFO and uses the baseline shallow temporal block, and the complete STME-GCN, as shown in Fig. 5.

Both actions mainly involve the bilateral upper limbs and hand joints and exhibit somewhat similar overall poses, but their motion processes differ substantially. Clapping is characterized by the periodic approach and separation of the two hands, whereas rubbing hands primarily consists of local reciprocating movements while the hands remain in continuous contact.

Compared with the variant without TME, the responses for clapping become more concentrated on the bilateral upper-limb joints and around the key temporal phases when the hands approach each other. More pronounced phase-wise fluctuations also appear in the frame-level response curve. In contrast, the responses for rubbing hands

remain continuously distributed over the principal moving joints for a longer duration, while the corresponding curve varies more smoothly. This pattern is consistent with its relatively small-amplitude but sustained motion. The distinct response distributions of the two actions indicate that the module does not simply increase the response intensity of all joints or frames. Instead, it highlights action-relevant joints and temporal intervals according to the specific motion pattern.

5 Conclusion

This work introduces STME-GCN, which combines temporally enriched channel-wise topology learning with explicit shallow-layer motion modeling for skeleton-based action recognition. The Topology Feature Optimization (TFO) Module incorporates global-state, local-motion, and phase-evolution cues into topology generation. The Temporal Motion Enhancement (TME) Module introduces inter-frame differences through a pose-motion collaborative pathway. With two-stream fusion, STME-GCN achieves top-1 accuracies of 92.6%/96.9% on NTU RGB+D 60 (X-Sub/X-View), 88.4%/90.0% on NTU RGB+D 120

(X-Sub/X-Set), and 95.3% on NW-UCLA. The consistent performance across datasets with different scales and evaluation protocols, together with the low joint-stream complexity, indicates a favorable balance between recognition accuracy and computational efficiency. Together with the visualization results, the ablation studies suggest that TFO improves action-dependent relation modeling, whereas TME emphasizes discriminative joints and temporal intervals. However, fixed phase partitioning may restrict adaptation to actions with irregular temporal progression, while skeleton-only inputs provide limited cues about interacting objects and fine-grained hand motions. Future work will explore adaptive temporal modeling and integrate appearance or object semantics to improve the recognition of complex actions.

Author contributions Author contributions will be provided after the author information is finalized.

Funding This work was supported by National Natural Science Foundation of China (U2568224 and 52372402), China.

Data availability The datasets analyzed in this study are publicly available from their original sources.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- [1] SUN Z, KE Q, RAHMANI H, et al. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3200–3225. doi: 10.1109/TPAMI.2022.3183112.
- [2] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press, 2018: 7444–7452. doi: 10.1609/aaai.v32i1.12328.
- [3] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2019: 12026–12035. doi: 10.1109/CVPR.2019.01230.
- [4] YE F, PU S, ZHONG Q, et al. Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition. In: *Proceedings of the 28th ACM International Conference on Multimedia*. New York, NY: ACM, 2020: 55–63. doi: 10.1145/3394171.3413941.
- [5] CHEN Y, ZHANG Z, YUAN C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway, NJ: IEEE, 2021: 13359–13368. doi: 10.1109/ICCV48922.2021.01311.
- [6] LIU Z, ZHANG H, CHEN Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2020: 143–152. doi: 10.1109/CVPR42600.2020.00022.
- [7] LI C, LI S, GAO Y, et al. Static graph convolution with learned temporal and channel-wise graph topology generation for skeleton-based action recognition. *Computer Vision and Image Understanding*, 2024, 244: 104012. doi: 10.1016/j.cviu.2024.104012.
- [8] ZHANG Y Q, PANG C, GENG P, et al. Multi-scale adaptive large kernel graph convolutional network for skeleton-based action recognition. *Journal of Computer Science and Technology*, 2025, 40(5): 1285–1300. doi: 10.1007/s11390-025-5287-1.
- [9] SHI L, ZHANG Y, CHENG J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 2020, 29: 9532–9545. doi: 10.1109/TIP.2020.3028207.

- [10] LEE J, LEE M, LEE D, et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway, NJ: IEEE, 2023: 10444–10453. doi: 10.1109/ICCV51070.2023.00958.
- [11] ZHOU Y, YAN X, CHENG Z Q, et al. BlockGCN: Redefine topology awareness for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2024: 2049–2058. doi: 10.1109/CVPR52733.2024.00200.
- [12] XIE J, MENG Y, ZHAO Y, et al. Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*. Washington, DC: AAAI Press, 2024: 6225–6233. doi: 10.1609/aaai.v38i6.28440.
- [13] LI J, YE P, XIA Y, et al. Disentangled adaptive multi-dimensional dynamic graph convolutional network for skeleton-based action recognition. *Neurocomputing*, 2025, 658: 131693. doi: 10.1016/j.neucom.2025.131693.
- [14] ZHU S, LI R, HU H. MD-STGCN: Dynamic topology and multi-scale temporal modeling for skeleton-based action recognition. *Information Sciences*, 2026, 752: 123572. doi: 10.1016/j.ins.2026.123572.
- [15] LI C, WEN C, SUN P, et al. TE-STGCN: Topology enhanced spatio-temporal graph convolutional network for skeleton-based action recognition. *Applied Soft Computing*, 2026, 193: 114835. doi: 10.1016/j.asoc.2026.114835.
- [16] DO J, KIM M. SkateFormer: Skeletal-temporal transformer for human action recognition. In: *Computer Vision–ECCV 2024*. Cham: Springer, 2025: 401–420. doi: 10.1007/978-3-031-72940-9_23.
- [17] ZHAO J, DAI J, ZHOU F, et al. Dual-path spatio-temporal Mamba for skeleton-based action recognition. *The Visual Computer*, 2025, 41(9): 6507–6519. doi: 10.1007/s00371-025-03950-5.
- [18] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2016: 1010–1019. doi: 10.1109/CVPR.2016.115.
- [19] LIU J, SHAHROUDY A, PEREZ M, et al. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(10): 2684–2701. doi: 10.1109/TPAMI.2019.2916873.
- [20] WANG J, NIE X, XIA Y, et al. Cross-view action modeling, learning, and recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2014: 2649–2656. doi: 10.1109/CVPR.2014.339.
- [21] DUAN H, WANG J, CHEN K, et al. PYSKL: Towards good practices for skeleton action recognition. In: *Proceedings of the 30th ACM International Conference on Multimedia*. New York: ACM, 2022: 7351–7354. doi: 10.1145/3503161.3548546.
- [22] SHAO Y, MAO L, YE L, et al. H²GCN: A hybrid hypergraph convolution network for skeleton-based action recognition. *Journal of King Saud University – Computer and Information Sciences*, 2024, 36(5): 102072. doi: 10.1016/j.jksuci.2024.102072.
- [23] JANG S, LEE Y, JEON T, et al. Dynamic partitioning graph convolutional network for skeleton-based action recognition. *ETRI Journal*, 2026, 48(1): 141–152. doi: 10.4218/etrij.2024-0598.
- [24] LU C, CHEN H, LI M, et al. Attention-guided and topology-enhanced shift graph convolutional network for skeleton-based action recognition. *Electronics*, 2024, 13(18): 3737. doi: 10.3390/electronics13183737.
- [25] CHENG K, ZHANG Y, HE X, et al. Skeleton-based action recognition with shift

graph convolutional network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2020: 183–192. doi: 10.1109/CVPR42600.2020.00026.

- [26] HUANG X, ZHOU H, WANG J, et al. Graph contrastive learning for skeleton-based action recognition. In: *Proceedings of the Eleventh International Conference on Learning Representations*. 2023.