

Human Behavior Recognition Based on Graph Neural Network and Cross-modal Collaborative Attention Mechanism

Caixian Ye

Guangzhou Xinhua University

Lijun Xu

xulijun1474@xhsysu.edu.cn

Guangzhou Xinhua University

Hao Qi

Guangzhou Xinhua University

Method Article

Keywords: Human behavior recognition, Gnet_CAM, Graph neural network, Dynamic spatiotemporal modeling, Collaborative attention mechanism

Posted Date: August 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10237948/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Discover Computing on October 15th, 2026. See the published version at <https://doi.org/10.1007/s10791-026-10690-0>.

Human Behavior Recognition Based on Graph Neural Network and Cross-modal Collaborative Attention Mechanism

Caixian Ye^{1,a}, Lijun Xu^{1,b,*} and Hao Qi^{2,c}

¹*School of Computing and Artificial Intelligence, Guangzhou Xinhua University, Guangzhou, China.*

²*School of Literature and Education, Guangzhou Xinhua University, Guangzhou, China.*

^a*Email: yecaixian@xhsysu.edu.cn*

^b*Email: xulijun1474@xhsysu.edu.cn*

^c*Email: qihao3723@xhsysu.edu.cn*

**Corresponding Author*

Abstract: By integrating complementary information from visual, skeletal, and audio sources to recognize human behavior in complex environments, this approach provides an effective means to enhance recognition robustness. However, existing methods still fall short in modeling dynamic spatiotemporal features and achieving adaptive cross-modal information collaboration, making stable recognition challenging in complex scenarios such as occlusion or varying lighting conditions. This paper proposes a novel hierarchical attention mechanism that operates both within and across modalities, integrating multimodal spatiotemporal feature data—including visual, audio, and skeletal information—extracted by graph neural networks, to develop a model for human behavior recognition (Gnet_CAM). This method first constructs a dynamic spatiotemporal graph to model both physical relationships based on human joints and non-physical relationships mediated by audio attention. Subsequently, it employs a hierarchical collaborative attention mechanism to achieve adaptive extraction and alignment fusion of spatiotemporal features both within and across modalities. On the NTU RGB+D120 dataset, this method achieved a recognition accuracy of 93.1%. In the self-generated dataset simulating harsh lighting conditions and severe occlusion, its performance exceeds that of the optimal baseline model by more than 8%. Ablation experiments demonstrated that removing either the audio attention edge or the cross-modal attention module results in a significant decline in model performance. Experimental results demonstrate that the proposed collaborative attention framework for cross-modal fusion modeling significantly enhances the robustness of behavior recognition in complex scenarios.

Keywords: Human behavior recognition; Gnet_CAM; Graph neural network; Dynamic spatiotemporal modeling; Collaborative attention mechanism

1. Intriduction

Human behavior recognition, as a fundamental topic in computer vision, holds core significance across numerous applications including intelligent surveillance, human-computer interaction, and virtual reality [1]. Traditional methods primarily rely on single modalities (such as RGB video or skeletal sequences) for analysis [2,3]. However, in complex real-world scenarios, single-modal data is susceptible to interference from environmental factors—including lighting variations, perspective differences, and occlusions—which significantly degrade recognition performance [4]. In recent years, the emergence of deep learning—particularly three-dimensional convolutional neural networks (3D CNNs)[5] and time-series modeling architectures such as LSTM[6] and Transformer[7]—has significantly enhanced behavioral recognition performance in video applications. Meanwhile, with the widespread adoption of low-cost depth sensing technologies, significant progress has been made in research based on human skeletal sequences, endowing these systems with enhanced robustness to

variations in viewing angles, appearance, and background [8–10]. Modeling the human body as a spatiotemporal graph and employing graph neural networks for feature learning has become the mainstream approach in this field.

However, single-modal perception has inherent limitations. Visual information is easily affected by environmental interference, while skeletal data may become incomplete due to sensor errors or severe occlusion. This has driven research toward multimodal fusion, which involves the synergistic utilization of heterogeneous data streams such as RGB images, skeletal data, and even audio. Researchers employed modal splicing or post-processing decision-making approaches to enhance performance, yet these methods often overlook the complex, dynamic spatiotemporal correlations between modalities [11–13]. Recently, attention mechanisms, particularly cross-modal attention, have been introduced to dynamically recalibrate and integrate features from different modalities [14,15]. However, designing a unified end-to-end framework for behavior recognition tasks that simultaneously achieves fine spatiotemporal dynamic modeling within modalities and adaptive, collaborative feature alignment across modalities remains a fundamental unresolved challenge.

To contextualize our research within the broader landscape of relevant methodologies, we note that recent studies have expanded human understanding based on video and attributes beyond traditional dual-flow and 3D-CNN frameworks. Ranjit et al. [16–17] proposed employing a Transformer-style multi-branched network for human attribute recognition, achieving robust attribute inference by jointly utilizing multiple feature sources and channel-aware attention mechanisms. Other studies have explored cross-modal and multi-scale fusion in image fusion or attention aggregation tasks [18,19], demonstrating the significant research value of integrating spatial, frequency, and temporal cues. Furthermore, recent studies have addressed pose-based recognition challenges in low/zero-sample behavior identification scenarios or complex conditions by employing attention relationship modeling or leveraging privileged semantic information [20–23]. Other studies focus on character attribute retrieval, anomaly detection, or robustness under low-light conditions [24,25]. These efforts collectively highlight the growing emphasis on multimodal fusion, attention-based adaptation, and generalization recognition that transcends fixed dataset scenarios. These motivations align perfectly with the algorithm we proposed, providing solid theoretical support.

Although both multimodal fusion and graph neural networks have achieved significant success, current cutting-edge approaches still face two interrelated fundamental limitations when dealing with complex scenarios. First, at the level of feature representation, most methods based on skeletal graphs rely on predefined, static anatomical connections of the human body [26]. This fixed topology cannot model non-physical semantic relationships that vary with actions and scene contexts—such as interactions between individuals based on visual focus or auditory attention—which are essential for understanding cooperative or adversarial behaviors. Secondly, at the integration mechanism level, existing cross-modal attention models typically focus solely on simple weighting or aggregation of features across modalities, lacking a hierarchical and collaborative framework to explicitly decouple and model dependencies across spatiotemporal dimensions before enabling cross-modal deep interaction [27]. This "coarse-grained" fusion approach may lead to information redundancy or confusion, making it impossible to perform robust inference when partial modal information is missing or damaged (e.g., due to severe occlusion causing loss of skeletal data).

In short, the current research gap lies in the absence of a unified framework that integrates dynamic spatiotemporal structure learning with fine-grained analysis. Specifically, an ideal solution should be capable of: (1) constructing a dynamic graph structure that simultaneously encodes both the

physical constraints within the human body and external, context-driven semantic relationships.(2) Design a collaborative attention mechanism that can both extract spatiotemporal features within each modality and guide selective, complementary alignment and fusion of features across different modalities.Overcoming these challenges is key to enhancing the robustness and generalization capability of behavior recognition systems in complex environments.

Based on the above analysis, this study proposes a novel and unified deep learning framework to address the dual challenges of dynamic spatiotemporal feature extraction and cross-modal collaborative fusion.We propose that by modeling human posture and scene context (particularly audio cues) as a dynamically evolving spatiotemporal graph and introducing a hierarchical collaborative attention mechanism, the model's performance in behavioral recognition and robustness in complex multimodal scenarios can be significantly enhanced.Therefore, the research objectives include: (1) Designing and implementing a dynamic multimodal graph neural network learning framework coupled with a hierarchical collaborative attention mechanism model (Gnet CAM), aimed at constructing a dynamic spatiotemporal graph model capable of uniformly encoding visual (RGB), skeletal, and audio information.At its core, the approach utilizes learnable "virtual attention edges" to link audio semantics with visual and skeletal nodes, overcoming the limitation of existing methods that rely solely on fixed human anatomical connections, and enabling explicit modeling of cross-modal non-physical associations. (2) The system evaluates and validates the effectiveness of the Gnet CAM model: it achieves leading recognition accuracy on the standard dataset (NTU RGB+D 120) and, through rigorous ablation experiments, quantitatively analyzes the specific contributions of the hierarchical collaborative attention mechanism and each modal component to overall performance, thereby demonstrating the necessity and validity of the core design at the data level. (3) Verify and demonstrate the practicality of the Gnet CAM model in real-world complex scenarios: By constructing and testing a custom dataset in an industrial environment with severe occlusion and lighting interference, this study empirically demonstrates the model's exceptional robustness under real-world challenges, providing a reliable technical solution for addressing behavioral understanding challenges in smart manufacturing and intelligent security domains.

The paper proposes a human behavior recognition method (the Gnet CAM model) based on graph neural networks and a hierarchical collaborative attention mechanism. Its key contributions include:

This study proposes an innovative multimodal human behavior recognition framework (Gnet_CAM model), which is an end-to-end architecture integrating dynamic spatiotemporal graphs with a hierarchical collaborative attention mechanism.

Systematic experiments have identified key design and scenario characteristics: This study not only demonstrates advanced performance (e.g., 93.1% accuracy on the NTU RGB+D 120 dataset), but also quantitatively confirms through rigorous ablation experiments that cross-modal attention (its removal resulted in a 4.7% performance decline) is essential for effective feature fusion.

This study has advanced the application of behavior recognition technology to real-world complex scenarios: it developed a comprehensive behavioral dataset comprising 350 videos under challenging conditions such as severe occlusion and uneven lighting, and demonstrated the proposed framework's exceptional robustness in such environments (93.9% accuracy, outperforming visual baselines by 8.4%).

The structure of the remainder of this paper is as follows: Section 2 provides a detailed exposition of the proposed method, covering the construction of the Gnet CAM model—including a dynamic spatiotemporal graph neural network, a hierarchical collaborative attention mechanism, and the design

of the loss function. Section 3 will describe the experimental setup and present quantitative and qualitative results across multiple publicly available and self-built datasets, including a comprehensive comparison with baseline methods and systematic ablation studies. Section 4 will provide an in-depth discussion on the significance of this study, its current limitations, and potential future research directions. Section 5 will summarize the entire manuscript.

2. Research method

This paper proposed human behavior recognition framework, the Gnet_CAM model, is illustrated in Figure 1. Implement the construction of dynamic spatiotemporal graph neural networks, develop cross-modal hierarchical collaborative attention mechanisms, and design loss functions. It primarily consists of four core modules: a multimodal feature encoder, a dynamic spatiotemporal graph construction module, a cross-modal hierarchical collaborative attention module, and a classification/identification prediction module.

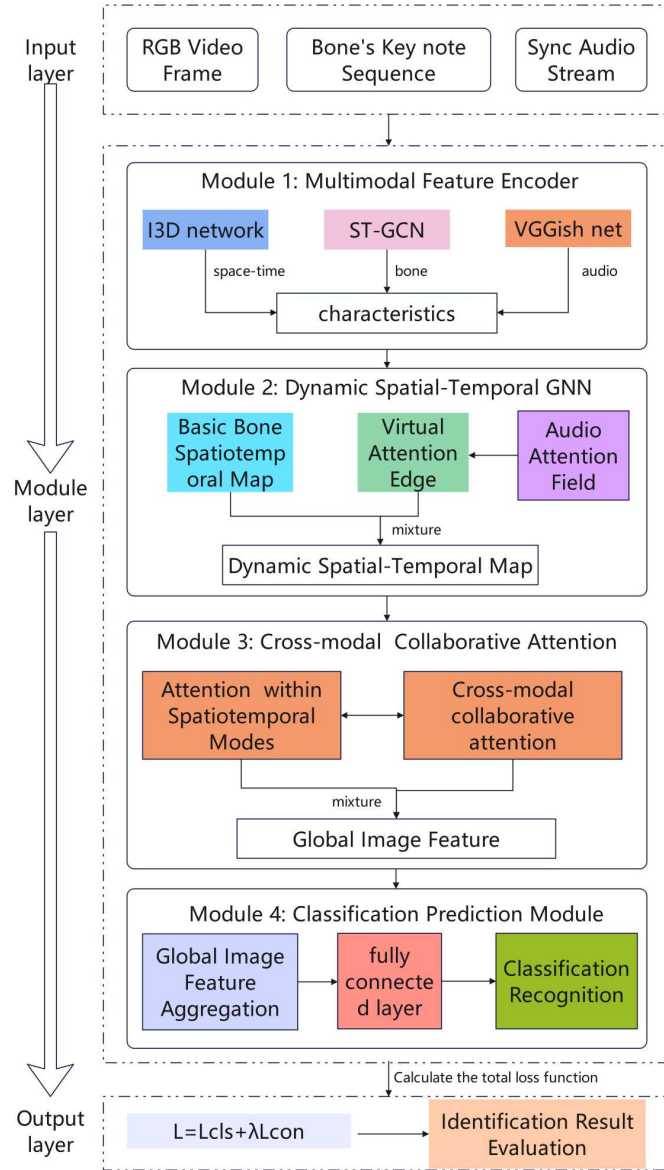


Figure 1. Multimodal Behavior Recognition Framework Gnet_CAM model

This research methodology implements the concept of constructing dynamic spatiotemporal graphs, demonstrating the integration process between a fundamental skeletal graph (representing

physical connections) and a virtual attention field driven by audio attention mechanisms. In the hierarchical collaborative attention mechanism, we design this module as a dual-path parallel architecture, emphasizing a modal refinement approach that performs spatiotemporal decoupling attention optimization separately for both the visual representation and skeletal representation. And achieves bidirectional, dynamic feature interaction and alignment among visual, skeletal, and audio modalities through the QKV (Query, Key, Value) mechanism.

We employ an end-to-end training pipeline that covers everything from raw data input to feature encoding, graph construction and fusion, and classification prediction. The complete closed-loop optimization process, achieved through a combined loss function (cross-entropy loss + spatiotemporal embedding comparison loss), demonstrates the model's end-to-end trainability.

2.1 Construction of Dynamic Spatiotemporal Graph Neural Networks

The dynamic spatiotemporal graph neural network integrates multimodal feature encoding with dynamic spatiotemporal graph construction. The Gnet_CAM model accepts three types of modal data as input: RGB video frames, human skeletal keypoint sequences, and synchronized audio streams. Therefore, the encoding method primarily relies on input visual features, skeletal features, and audio features. For visual feature encoding, a pre-trained I3D network on the Kinetics dataset is employed to extract spatiotemporal features from RGB videos (<https://github.com/deepmind/kinetics-i3d>). Bone feature encoding employs the NTU RGB+D 120 standard dataset to model the human skeleton as a spatiotemporal graph $G = (V, E)$, where nodes V represent joints, and edges E include both physical connections between body parts (spatial edges) and connections of the same joint across consecutive frames (temporal edges). The spatiotemporal graph convolutional network (ST-GCN) developed by Lei Jianyun et al. was employed to extract skeletal motion features [28]. Audio feature encoding employs a VGGish network to extract a 128-dimensional audio feature vector and map it into a representation with the same dimension as visual features [29].

Following multimodal feature encoding, we implemented a dynamic spatiotemporal graph extension to model cross-modal non-physical correlations by introducing audio-related "virtual attention edges" onto the baseline skeleton graph. The generation mechanism of the virtual attention edge involves projecting heterogeneous features into a shared space to enable cross-modal comparison, calculating attention scores to obtain normalized correlation strengths between audio signals and each node, generating weighted contextual representations that form a global "audio attention field," and finally computing the affinity between each node and the audio field as the generation strength of the virtual edge, thereby producing the virtual attention edge. For example, establish a correlation between the head joints and the direction of the sound source. The steps are shown in Table 1:

Table 1. Steps for generating "Virtual Attention Edges"

Steps	objectives	Input	Core Formulas	Output
1.Feature Projection and Unified Representation	Map heterogeneous audio data and node features into the same semantic space d_{model} to calculate their correlations.	Global audio feature a and the node feature set V .	Audio Search: $Q=aW_q+b_q$; Node Key: $K=VW_k+b_k$; Node value : $V_{proj}=VW_v+b_v$ Where W_q , W_k , and W_v are learnable projection matrices, and b_q , b_k , and b_v are biases.	Unified Dimension Query Q , Key K , Value V_{proj} .
2.Calculate the cross-modal attention score	Measures the correlation between audio and each node, then normalizes it to a probability distribution.	Query Q and key K .	Use Scale Product Attention: $s=\frac{QK^T}{\sqrt{d_{model}}}$ Perform Softmax normalization on column s : $\alpha_i=\frac{\exp(s_i)}{\sum_{j=1}^N \exp(s_j)}$ $\alpha=[\alpha_1,\alpha_2,...,\alpha_N]$, α_i represents the attention strength of audio to node i .	The attention weight vector α satisfies. $\sum_i \alpha_{i=1}$
3.Generate Audio Attention Field	Use attention weights as coefficients to aggregate the projection values of all nodes, creating a global audio-conditioned context representation.	Weight α and projection value V_{proj}	Weighted Sum: $C=\sum_{i=1}^N \alpha_i \cdot V_{proj}^{(i)}$ Where $V_{proj}^{(i)}$ denotes the projection feature of the i -th node.	The Audio Attention Field C encodes the visual/skeletal context information relevant to the audio.
4.Generate virtual edges from fields	Calculate the connection strength of virtual edges based on each node's compatibility with the audio attention field.	The characteristics of the original node V_i and the audio attention field C .	For each node i : 1. Concatenate node features V_i and C : $Z_i=[V_i\ C]$ 2. Using a small multi-layer perceptron (MLP) with a sigmoid activation function to output scalar intensity values: $e_i=\sigma(MLP(Z_i))$ Where σ 为 Sigmoid function.	Each node corresponds to a virtual edge strength, representing the probability or weight of establishing a connection between that node and the audio source.

The core concept of this mechanism is that the audio signal acts as an active query agent, which uses an attention mechanism to "examine" the entire visual/skeletal graph, identifies the spatial location (node) most relevant to the current sound, quantifies this relevance using the attention weight α , and ultimately guides the formation of virtual edges through the attention field. The core lies in a cross-modal attention layer. This layer enables audio features to serve as "queries" that capture and aggregate relevant information from visual/skeleton graph nodes (acting as "keys" and "values"), producing an "audio attention field" that guides the generation of virtual edges. The basic skeletal spatiotemporal graph is fused with this virtual attention edge to generate a processed dynamic spatiotemporal graph, from which the feature Z_i of the audio attention field C integrated into the skeletal graph node at layer i is obtained, thereby providing an effective dataset for the subsequent section.

2.2 Hierarchical Collaborative Attention Mechanism

The core mechanism of the Gnet CAM model is a hierarchical collaborative attention mechanism, designed to achieve both intra-modal feature optimization and cross-modal feature synergy.

- Optimize attention within the spatiotemporal modality

The attention module within the spatiotemporal modality serves as the fundamental unit for feature purification in a multimodal behavior recognition framework, with its primary objective being to perform precise feature recalibration separately on the visual modality (RGB feature map) and the skeletal modality (skeleton graph). The design of this module is based on a fundamental insight: the semantics of an action are jointly defined by the spatial coordination of different body parts and the temporal evolution of their states. However, the spatially significant joints and temporally critical frames are not fixed but highly dependent on the specific action category and execution context. Therefore, we adopt a spatiotemporal decoupling strategy by breaking down the complex spatiotemporal attention mechanism into two relatively independent and more manageable sub-problems—space and time—allowing us to focus separately on "where is important" and "when is important."

Based on the aforementioned skeletal feature encoding, we model the human skeletal structure as a spatiotemporal graph $G = (V, E)$ for skeletal representation. The set V contains all human joint points in each frame, while the edge set E includes two types of connections: (1) spatial edges: edges formed within the same frame based on natural human movements (e.g., an elbow connecting the wrist to the shoulder); (2) temporal edges: edges formed between adjacent frames through connections at the same joint point.

The features of the audio attention field C are integrated from the skeletal graph nodes at layer i in the previous section $Z_i \in \mathbb{R}^{T \times N \times C}$ (T As frames, N is Node's number, C is the feature channel's number), Attention within the spatiotemporal modality is processed through parallel spatial attention branches and temporal attention branches, with its core computational workflow illustrated in Figure 2.

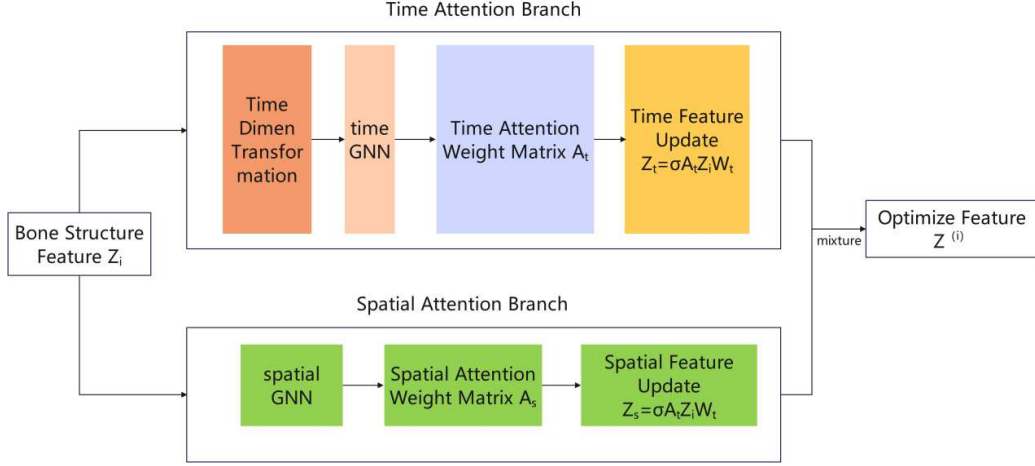


Figure 2. Attention calculation workflow within spatiotemporal modalities

The spatial attention branch is designed to learn the relative importance of different body joints for the current action within each frame. First, we utilize a learnable spatial adjacency matrix A_s to initialize the connection strength between joints. This matrix differs from the standard physical connection adjacency matrix in that its element a_{ij} can be optimized using gradient descent, enabling it to adaptively capture joint associations related to action semantics and physical connections. For example, during the "throwing" movement, the hands and feet may exhibit strong correlation in maintaining balance. The spatial attention weight is computed using the self-attention mechanism between features, as shown in Formula (1):

$$A_s^{(l)} = \text{Softmax}\left(\frac{(Z^{(l)} W_Q^s)(Z^{(l)} W_K^s)^T}{\sqrt{d_k}}\right) \quad (1)$$

There W_Q^s , W_K^s are learnable projections matrix used in the spatial dimension to generate queries and keys. d_k is the flowering factor. Subsequently, the spatial features are updated using this weight as shown in Formula (2):

$$Z_s^{(l+1)} = \sigma(A_s^{(l)} Z^{(l)} W_V^s) \quad (2)$$

There W_V^s is the value projection matrix, and σ is the nonlinear activation function (we use ReLU). This operation enhances the feature response of key joints (e.g., the wrist during a "waving" motion) while suppressing interference from irrelevant or noisy joints.

The additional temporal attention pathway is designed to capture the evolution patterns of actions across time and emphasizes key frames that play a decisive role in recognition. First, we transpose the features to construct a "time graph" along the temporal dimension, where nodes represent the states of the same joint at different time points. The calculation method for time attention weight is similar to that for spatial attention, but it operates along the temporal dimension as shown in Formula (3), and the corresponding feature update is given by Formula (4):

$$A_t^{(l)} = \text{Softmax}\left(\frac{(Z^{(l)} W_Q^t)(Z^{(l)} W_K^t)^T}{\sqrt{d_k}}\right) \quad (3)$$

$$Z_t^{(l+1)} = \sigma(A_t^{(l)} Z^{(l)} W_V^t) \quad (4)$$

This mechanism enables the model to automatically focus on peak or transition frames of actions (e.g., the instant from standing to sitting in the "sit down" action) while smoothing out the effects of preparatory or redundant frames.

With both branches prepared, we merged the updated features by combining those optimized through spatial and temporal attention fields respectively. The updated output feature $Z^{(l+1)}$ for layer l is obtained as shown in Formula (5):

$$Z^{(l+1)} = \text{LayerNorm}(Z_s^{(l+1)} + Z_t^{(l+1)}) \quad (5)$$

This residual connection combined with LayerNorm ensures training stability and enables efficient information fusion. $Z^{(l+1)}$. For the visual modality (RGB feature map), we employ the same spatiotemporal decoupled attention architecture and apply it to the grid feature map extracted by CNN. Spatial attention operates on the spatial locations within feature maps, identifying which spatial regions in the scene are relevant to behavior; temporal attention acts on frame sequences, following a principle consistent with skeletal modal representation.

Thus, the optimized spatiotemporal attention module decouples and concurrently models dynamic importance across spatial and temporal dimensions, generating more discriminative feature representations for each modality. This lays a solid foundation for subsequent cross-modal collaborative attention, ensuring that the features within each module are thoroughly refined and rich in semantic information before cross-modal interaction.

- Cross-modal collaborative attention

Although the modal features optimized through intra-modal attention possess rich internal spatiotemporal context, they remain within a heterogeneous feature space. The core objective of the cross-modal collaborative attention module is to establish a dynamic, bidirectional, and iterative "dialogue" mechanism across the three modalities—visual, skeletal, and audio. This mechanism enables features from any modality to serve as "queries" and proactively retrieve relevant information from key-value pairs in other modalities, thereby achieving cross-modal semantic alignment and feature complementarity. This iterative collaborative approach aims to simulate the integration process of human multi-sensory perception—for instance, when hearing applause, the system instinctively focuses on the hand region in the visual image, thereby achieving a paradigm shift from "feature stitching" to "feature symbiosis."

This module operates in a cyclic iterative manner. In each iteration cycle, modal features interact through a shared cross-modal attention layer, with its output serving as input for the subsequent round, thereby enabling deep integration of cross-modal information. The core computational process of each iteration is shown in Figure 3:

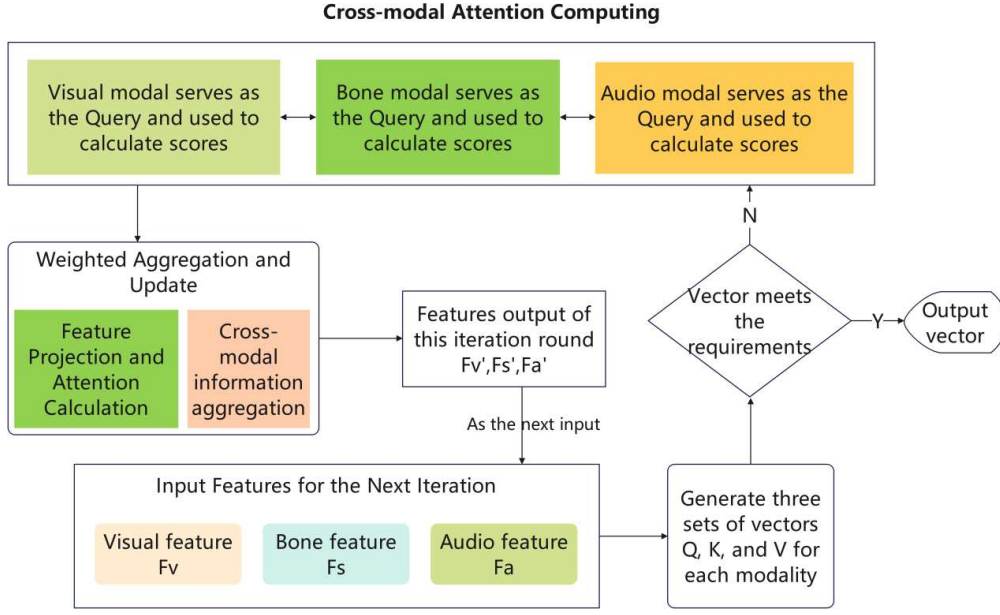


Figure 3. Attention calculation workflow within spatiotemporal modalities

The key to cross-modal attention computation lies in performing calculations and updates for the following two components:

1) Feature Projection and Attention Calculation

Assuming at the t -th iteration, the features of the visual, skeletal, and audio modalities are as F_v^t , F_s^t and F_a^t . First, Formula (6) generates queries, keys, and value projections for each modality. Taking the query of visual modal information using audio modal as an example, Formula (7) defines the attention score as follows:

$$Q_m^t = F_m^t W_Q^m, K_m^t = F_m^t W_K^m, V_m^t = F_m^t W_V^m, m \in \{v, s, a\} \quad (6)$$

$$A_{a \rightarrow v} = \text{Softmax}\left(\frac{Q_a^t (W_v^t)^T}{\sqrt{d_k}}\right) \quad (7)$$

Here, W_Q^m, W_K^m, W_V^m are mode-specific learnable projection matrices. Formula (7) quantifies the correlation between each audio feature unit and all visual feature units. For example, an audio feature representing "door hinge creaking" may exhibit a high attention score when correlated with the spatial positions of "door" and "hand" in visual features.

2) Cross-modal information aggregation and feature updating

Using the computation-derived attention weights, the values are weighted and summed to aggregate information from other modalities into the current modality, as shown in Formula (8):

$$\tilde{V}_a^t = A_{a \rightarrow v} V_v^t + A_{a \rightarrow s} V_s^t \quad (8)$$

Here $\tilde{V}_a^t$ refers to cross-modal contextual information aggregated from visual and skeletal modalities to the audio modality. To preserve the original information of the modality while integrating new information, we employ a gated fusion mechanism (as shown in Formula (9)) to update the audio features:

$$G_a^t = \sigma(F_a^t W_g^a + \tilde{V}_a^t U_g^a + b_g^a) \quad (9)$$

$$F_a^{t+1} = \text{LayerNorm}(G_a^t \odot \tilde{V}_a^t + (1 - G_a^t) \odot (F_a^t W_r^a))$$

Here, σ denotes the Sigmoid function. W_g^a, U_g^a, W_r^a are learnable parameters. $\odot$ indicates element-wise multiplication. The gating vector G_a^t dynamically determines how much new information

is absorbed from other modalities. The feature updates for both visual and skeletal modalities are performed in exactly the same manner, enabling bidirectional information flow. For instance, when updating visual features, precise joint motion information can be simultaneously obtained from the skeletal modality, while contextual scene sounds can be acquired from the audio modality.

3) ITERATIVE OPTIMIZATION

The above process constitutes a complete iteration. We use the updated features F_v^{t+1} , F_s^{t+1} , F_a^{t+1} as inputs for the next iteration ($t+1 \rightarrow t+2$). Typically, after 2–3 iterations, the modal features achieve a stable collaborative state. Finally, we combine the modal features generated during the last iteration. The globally generated cross-modal collaborative representation F_{fusion} is then fed into the downstream classification identifier.

During the iteration process, each modality both contributes and consumes information, ultimately aligning all modalities' features semantically toward a shared behavioral concept, thereby laying a solid foundation for precise behavior recognition.

2.3 Classification Feature Recognition and Loss Function Design

This section details how the model transforms learned collaborative features into final classification results, and how a carefully designed loss function guides and optimizes the entire multimodal learning process. This process aims to ensure that the model not only makes accurate predictions but also learns essential, robust semantic representations of actions.

● 2.3.1 Classification Feature Recognition

Through the deep processing involving the aforementioned multi-layer graph attention network and iterative cross-modal attention module, we have obtained highly refined and synergistic multimodal features. For the skeletal modality, the final node feature matrix $Z(L)$ (where L denotes the total number of layers) was obtained; corresponding optimized features were also derived for the visual and audio modalities. To generate a global representation for classification, we employ a hierarchical aggregation strategy:

1) Node-level Pooling: First, apply maximum pooling (Max Pooling) to each node's feature sequence along the time dimension T to capture its most discriminative states throughout the entire action cycle. The purpose of node-level aggregation is to eliminate variability across time dimensions by compressing the dynamic temporal features of each key node into a static, highly representative feature vector.

The final feature tensor $Z(L)$ resulting from the processing of skeletal modal data through an L -layer network. Here, T denotes the number of time frames, N represents the number of joints, and C stands for the number of feature channels. This step aims to eliminate variability across time dimensions by compressing the temporal features of each node into a highly discriminative static feature vector. The standard procedure involves performing maximum pooling on each node n along the time dimension T . The feature is compressed into Z' as shown in Formula (10):

$$Z'_n = \max_{t \in [1, T]} Z^{(L)}[t, n, :], \forall n \in [1, N] \quad (10)$$

Here, Z'_n represents the aggregated feature of node n . By stacking the features of all nodes, we obtain the aggregated feature matrix at the node level: $Z'_n = [Z'_1; Z'_2; \dots; Z'_N]$.

Graph-level Pooling: Subsequently, Z' undergoes both average pooling and maximum pooling along the spatial dimension N (i.e., across all nodes), with the results of both methods being concatenated to form the global feature vector f_m for this modality. The purpose of level-based aggregation is to consolidate spatially dispersed node features into a global representation of the entire

human motion.

By inputting the aggregated feature matrix Z' at the node level, we calculate the average values across all feature channels along the node dimension N , capturing the common response patterns of all joints. As shown in Formula (11):

$$f_{avg} = \frac{1}{N} \sum_{n=1}^N Z_n \quad (11)$$

Then, maximum pooling is applied by extracting the maximum value across all feature channels along the node dimension N to capture the most discriminative local significant responses, as shown in Formula (12):

$$f_{max} = \max_{t \in [1, N]} Z_n \quad (12)$$

Finally, feature concatenation is performed by merging the two pooled vectors to form the global feature vector f_m for this modality. Extending this process, the global feature vectors for the visual, skeletal, and audio modalities are denoted as f_v , f_s , f_a respectively.

2) Multimodal fusion: By concatenating the global feature vectors from visual, skeletal, and audio modalities, a unified fusion feature is generated, as expressed in formula (13):

$$f_{fusion} = \text{Concat}(f_v, f_s, f_a) \quad (13)$$

The resulting f_{fusion} represents a fusion feature that integrates multimodal spatiotemporal information, retains the unique discriminative features of each modality, and provides a foundation for subsequent interaction and reweighting.

3) Classifier and Prediction Output

Finally, f_{fusion} is fed into a lightweight classification network to predict behavioral categories. The classifier consists of a two-layer fully connected network, with the computational process described in Formula (14):

$$\begin{aligned} h &= \text{Dropout}(\text{ReLU}(W_1 f_{fusion} + b_1)) \\ y &= \text{Softmax}(W_2 h + b_2) \end{aligned} \quad (14)$$

There, W_1 is the weight matrix of the first fully connected layer in the classifier, W_2 is the weight matrix of the fully connected layer in the output layer. b_1 are the parameters of the first layer, and b_2 are the parameters of the second layer. $\text{ReLU}()$ is an activation function, and $\text{Dropout}()$ prevents overfitting. y represents the category probability distribution of the model output, where its k -th element denotes the predicted probability that a sample belongs to class k .

W_1 is responsible for feature transformation and compression, while W_2 generates the final classification score. Together with their corresponding bias parameters b_1 and b_2 , these parameters form the core of the classifier. During model training, they are iteratively updated through backpropagation and optimization algorithms to minimize the loss function, enabling the model to accurately identify behavioral categories from fused features.

● 2.3.2 Loss Function Design

To guide the model's learning, we designed a composite loss function composed of a primary loss for supervised classification and an auxiliary contrast loss with self-supervision.

1) Main Loss: Standard Cross-Entropy Loss

The primary loss function employs a standard cross-entropy loss function, designed to minimize the discrepancy between the model's predicted distribution $\hat{y}$ and the true labels y (in one-hot encoding), thereby directly optimizing classification accuracy.

$$\mathcal{L}_{cls} = - \sum_{c=1}^C y_c \log \hat{y}_c \quad (15)$$

Where C represents the total number of behavior categories.

2) Auxiliary Loss: Spatial-Temporal Embedding Comparison Loss

To enable the model to learn more fundamental and robust action representations, relying solely on classification loss is insufficient; we introduce a spatio-temporal embedding contrastive loss. The core idea is: within the feature embedding space, it brings together (positive sample pairs) feature representations that belong to the same action semantics but originate from different modalities or time slices, while separating (negative sample pairs) feature representations of distinct actions.

For a training batch, given an anchor sample feature Z_i , We constructed the positive and negative sample sets according to the following rules, as shown in Table 2:

Table 2. Positive and Negative Sample Sets

Sample Type	Create Rules	Rule Description	target
Positive Sample	Cross-modal positive sample	Global features of different modalities within the same video sample at the same time slice.	Close-up feature
	Cross-time positive sample	Features of adjacent time slices within the same video sample and the same modality.	Close-up feature
Negative sample	Negative samples within the batch	Within the same batch, all other distinct video samples share identical features regardless of their modality or time slice.	Remote Push Feature

Based on the loss framework, the comparison loss for anchor point Z_i is calculated as shown in Formula (16).

$$\mathcal{L}_{con}^{(i)} = - \log \frac{\sum_p \exp(\text{sim}(Z_i, Z_p)/\tau)}{\sum_p \exp(\text{sim}(Z_i, Z_p)/\tau) + \sum_n \exp(\text{sim}(Z_i, Z_n)/\tau)} \quad (16)$$

Here $p \in P(i)$, $n \in N(i)$, $P(i)$ and $N(i)$ are the positive and negative sample sets of anchor point i , respectively. $\text{sim}(Z_i, Z_*)$ denotes the cosine similarity; τ is a heat hyperparameter that regulates the emphasis on difficult negative samples. Therefore, the overall batch comparison loss is the average of all anchor point losses, as shown in Formula 17.

$$\mathcal{L}_{con} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{con}^{(i)} \quad (17)$$

3) Total Loss Function

Finally, the total loss for model training is the weighted sum of the two aforementioned losses:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{con} \quad (18)$$

Here, λ is a hyperparameter used to balance the contributions of the two loss terms. Through joint optimization of $\mathcal{L}_{total}$ the model is not only guided to make accurate classification decisions but also constrained to learn a well-structured feature space. Within this space, the multi-modal representations of identical actions are highly clustered, while those of different actions remain distinct from each other, significantly enhancing the model's robustness to noise and modal missingness as well as its generalization capability.

3. Experiment

3.1 Experimental Preparation

This dataset employed in this study primarily consists of a self-developed industrial complex scenario dataset and the open-source dataset NTU RGB+D 120 (<https://rosel.ntu.edu.sg/dataset/actionRecognition>). We selected the NTU RGB+D 120 skeletal action dataset, which contains 15 categories of actions and 16,000 video samples, for model training.

To evaluate the model's performance under real-world harsh conditions, we conducted validation experiments on 350 videos from our self-collected dataset of complex scenarios, covering 15 common movement types (standing, sitting, squatting, lying down, etc.). These videos exhibited challenges such as uneven lighting, dust interference, and severe occlusion. After determining the dataset, we conducted comparative experiments with current mainstream methods—including skeleton-based 2s-AGCN [28]; vision-based I3D [29]; and audio-based VGGish [30] as baseline approaches—all evaluated through systematic ablation studies on our self-developed industrial complex scene dataset. We adopt Top-1 classification accuracy as the primary evaluation metric.

Experimental Environment Description: The system employs an Intel Core i5-14600KF CPU, a NVIDIA RTX 5060 Ti graphics card with 16 GB of VRAM, Windows 11 Professional operating system, 32 GB of DDR4 3600 MHz RAM, and a 2 TB solid-state drive. Python 3.8 was used for programming.

3.2 Results and Analysis

In comparison with current mainstream methods, we conducted comparative experiments; additionally, we performed systematic ablation studies on our self-developed dataset of industrial complex scenarios. The experimental results are illustrated in Figure 4:

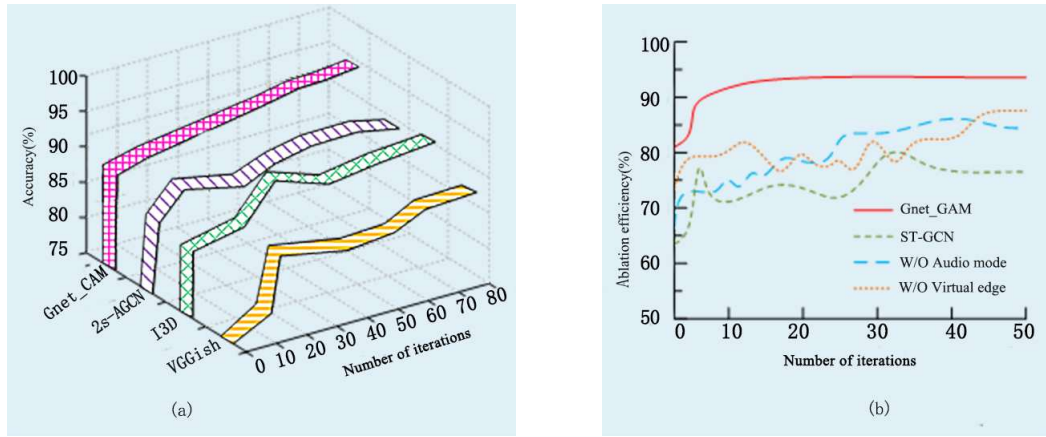


Figure 4. Comparison of validation results

(a) Compared with mainstream methods (b) Comparison of ablation experiments

Compared with other methods, we achieved at least a 4.6% improvement and demonstrated greater stability during iteration than other models, indicating that the Gnet CAM model benefits from more refined dynamic graph modeling and a collaborative attention mechanism.

Table 3 presents the comparison results on the NTU RGB+D 120 dataset. The proposed method achieves an accuracy of 93.1%, significantly outperforming all baseline models, demonstrating the effectiveness of our multimodal fusion approach.

Table 3. Comparison of recognition accuracy on the NTU RGB+D 120 dataset

Method	modality	Accurary (%)	illustration
2s-AGCN	bones	88.6	Adaptive graph convolutional network with a strong baseline for skeletal modalities
I3D	RGB	84.2	A dual-flow expansion 3D convolutional network based on Kinetics pre-training
VGCish	audio	76.8	Deep audio feature networks pre-trained using large-scale audio datasets
Gnet_CAM	Multimodal (Visual + Bone + Audio)	93.1	Our methods

To validate the contributions of each module and the effectiveness of the proposed methodology, we conducted a systematic ablation study on our self-developed dataset of industrial complex scenarios, with results presented in Table 4. The advantages of this method are even more pronounced, achieving an accuracy rate of 93.9%, which is 6.2% higher than that of the second-best approach. This fully demonstrates the exceptional robustness of our method in addressing real-world challenges (such as occlusion and lighting variations).

Table 4 Results of ablation experiments on the industrial complex scenario dataset

Model Variant	Modification Description	Accuracy (%)	Decrease Amount (Percentage)
Gnet_CAM	-	93.9	-
W/O Virtual Attention Edge	Keep only physical connection edges and remove virtual edges generated by the audio attention field.	87.7	↓6.2
W/O Audio Mode	Use RGB and bone modes.	86.2	↓7.7
ST-GCN	Use only bone data as a strong baseline for a single modality.	81.2	↓12.7
* w/o means “without”.			

Using the complete model as a benchmark, comparisons were made from core modules to individual modalities. Table 3 clearly illustrates the gradient of performance impact ranging from "failures in core model design" to "lack of information sources". Removing the cross-modal attention edge resulted in a 6.2% performance degradation, conclusively demonstrating that the collaborative attention mechanism proposed in this paper is pivotal for achieving effective multimodal fusion—its impact far surpasses that of simple feature concatenation. This demonstrates that the proposed method effectively addresses visual interference issues and achieves stable person recognition even in complex scenarios with occlusions and varying lighting conditions.

4. Discuss

The multimodal fusion framework proposed in this study demonstrates outstanding performance in both standard and complex scenarios. This discussion aims to conduct an in-depth analysis of the experimental results, elucidate the underlying advantages by comparing them with existing studies and performing internal ablation analyses, while objectively assessing their current limitations.

This method achieves a 93.1% accuracy rate on the NTU TGB+D 120 dataset, and its core advantage does not stem from extreme optimization for a single modality. Compared to recent studies that solely focused on skeletal data representation during the same period, this work establishes a more versatile and robust fusion architecture. This value is fully demonstrated in our self-built industrial complex scenario dataset: under severe visual degradation conditions, the model achieved an accuracy rate of 93.9%. This marks a pivotal step in our research, shifting from pursuing "laboratory precision" to addressing "robustness in real-world environments" —by leveraging cross-modal complementarity to mitigate the risks associated with single-source signal failures, a critical advantage for practical deployment.

The systematic ablation experiments (Table 4) revealed the underlying mechanisms of performance advantages at the data level. First, removal of the cross-modal collaborative attention module resulted in a significant performance degradation (-6.2%), directly demonstrating its critical role in achieving deep information fusion—far surpassing that of simple feature concatenation. Secondly, removing the virtual attention edge resulted in performance loss comparable to discarding the entire audio modality, strongly demonstrating that this design effectively links audio semantics (e.g., specific device sounds) to the core pathways of visual space and human joints rather than serving as a redundant structure. These findings collectively validate the effectiveness of this framework's design principles of "dynamic association modeling" and "fine-grained collaboration."

More intriguingly, the ablation experiment prompted a reevaluation of the relative importance of different modalities in complex scenarios. In complex scenario datasets, models utilizing only skeletal modalities achieve relatively low accuracy (81.2%), a finding also reported by Li Hang's team [31] in their study. Although the audio modality contributes relatively modestly overall, its unique cross-scenario semantics (e.g., tool operation sounds and voice commands) provides contextual information unattainable through visual or skeletal data, thereby enhancing the system's scene understanding and fault tolerance capabilities. This finding aligns with the conclusions of Saghir Alfasly et al. [32].

Of course, our research also has certain limitations. The current framework imposes stringent requirements on the synchronization and integrity of multimodal data, yet practical applications often encounter issues such as missing modalities or asynchronous processing. Furthermore, the construction of dynamic graphs and the multi-layer attention mechanism entail significant computational complexity, which may limit their deployment on resource-constrained devices. Future work will focus on exploring asynchronous multimodal fusion strategies and lightweight model designs, while expanding the framework to encompass a broader range of open-world scenarios and diverse perception modalities (such as thermal imaging and radar point clouds), thereby advancing its evolution into a more versatile and practical environmental intelligence system.

5. Conclusion

This study proposes a multimodal behavior recognition method based on dynamic spatiotemporal graphs and cross-modal collaborative attention mechanisms. Its core lies in leveraging dynamic graph

structures (which integrate physical connections with audio-driven virtual edges) and hierarchical attention mechanisms (first optimizing within each modality, then coordinating across modalities) to address information fusion and robust perception challenges in complex scenarios. On both publicly available benchmarks and self-developed industrial datasets, this method achieved an advanced accuracy rate of 93.9%. Ablation experiments demonstrated that the cross-modal attention module is pivotal to performance (a 6.2% decline occurs upon removal), while virtual attention edges are essential for audio information processing (their removal yields effects equivalent to eliminating the entire audio modality), fully validating the framework's robustness in addressing real-world challenges. Work provides an effective approach for understanding behavior in complex environments, and its core concept offers valuable insights for broader cross-modal learning tasks. Current limitations stem from data synchronization dependencies and model complexity; future efforts will focus on asynchronous integration and lightweight design to facilitate practical deployment.

Funding

The Key Scientific Research Project of Guangzhou Xinhua University (No.2025KYZDZK02) , Guangdong Intangible Cultural Heritage Research Team of the Guangdong-Hong Kong-Macao Greater Bay Area Project(No.2024KYCXTD01) and 2025 Key Research Platforms and Projects of Guangdong Province's Regular Higher Education Institutions(No.2025WTSCX150).

Declaration of Conflicting Interests

The author(s) declare(s) that there are no conflicts of interest regarding the publication of this paper

Data availability statement

Data can be obtained from the corresponding author on a reasonable request.

Declaration of Conflicting Interests

The author(s) declare(s) that there are no conflicts of interest regarding the publication of this paper

Ethical Statement: Not Applicable

Consent to Participate: not applicable

Consent to Publish: not applicable

Clinical trial number: not applicable

Author Contributions

Caixian Ye: Original draft preparation, funding acquisition and editing data curation.

Lijun Xu: Supervision, investigation, methodology and software.

Hao Qi: Resources, revision and funding acquisition.

Reference

- [1] Zhao H, Liu J, Wang W J, et al. Research on human behavior recognition in video based on 3DCCA[J]. MULTIMEDIA TOOLS AND APPLICATIONS, 2023:20251-20268. <https://doi.10.1007/s11042-023-14355-8>.

- [2] Wang T H, Cao J T , J i X F, et al. A new human motion recognition framework based on the fusion of joint data and RGB video[[C]. PROCEEDING OF THE 36TH CHINESE CONTROL AND DECISION CONFERENCE, 2024:3740-3745. <https://doi.org/10.1109/CCDC62350.2024.10587777>
- [3] Chen H, Pan Y K, Wang C W, et al. An optimization method of human skeleton keyframes selection for action recognition[J]. COMPLEX & INTELLIGENT SYSTEMS, 2024: 4659-4673.
- [4] Bhowmik A, Kumar S, and Bhat N. Evolution of automatic visual description techniques-a methodological survey. Multimedia Tools and Applications. 2021, 80(18): 28015-28059. <https://doi.org/10.1007/s11042-021-10964-3>
- [5] Wu H B, Ma X and Li Y B. Spatiotemporal Multimodal Learning with 3D CNNs for Video Action Recognition[J]. IEEE TRANCTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2022:1250-1261. <https://doi.org/10.1109/TCSVT.2021.3077512>
- [6] Wang JQ, Liu K and Gong Y. Vehicle Position Prediction Using Particle Filtering Based on 3D CNN-LSTM Model. IEEE Transactions on Mobile Computing. 2023, 23(4): 2992-3004. <https://doi.org/10.1109/TMC.2023.3267853>
- [7] Chen N. A spatiotemporal separation-based reverse Transformer action recognition model. Intelligent Internet of Things Technology. 2025, 57(01): 92-95. (in Chinese)
- [8] Qian HF, Yi JP, Fu YH. Review of human action recognition based on deep learning. Journal of Frontiers of Computer Science and Technology. 2021, 15(3): 438-455. (in Chinese)
- [9] Bhowmik A, Kumar S, and Bhat N. Evolution of automatic visual description techniques-a methodological survey. Multimedia Tools and Applications. 2021, 80(18): 28015-28059. <https://doi.org/10.1007/s11042-021-10964-3>
- [10] Chen LY, et al. Review of Image Classification Algorithms Based on Convolutional Neural Networks. Remote Sensing. 2022, 13. <https://doi.org/10.3390/rs13224712>
- [11] Zheng D C, Chen C W, Yu J Z, et al. Human behavior recognition algorithm based on multi-modal sensor data fusion[J].JOURNAL OF ADVANCED COMPUTATIONAL INTELLIGENCE AND INTELLIGENT INFORMATICS,2025:287-305.<https://doi.org/10.20965/jaciii.2025.p0287>
- [12] Xie Hairong, Hu Chaobin, Liang Guohua, et al. Intelligent assessment of psychological stress in college students: Design and validation of a multimodal model integrating text and imagery [J]. Journal of Southern Medical University.2025: 2504-2510.<https://doi.org/10.12122/j.issn.1673-4254.2025.11.23>(in Chinese)
- [13] Liu Yujie, Wang Yiwen. A micro-attitude recognition method based on multimodal collaborative enhancement [J]. Computer Engineering,2025 : <https://doi.org/10.19678/j.issn.1000-3428.0252355>.(in Chinese)
- [14] Dong Yangyang, Yang Kongjing, Feng Jiachen. A method for cross-modal feature integration-based interference identification based on the collaborative attention mechanism [J]. Journal of Military Engineering.2025.[https:// DOI.10.12382/bgxb.2025.0725](https://doi.org/10.12382/bgxb.2025.0725) .(in Chinese)
- [15] Bui D C, Le T V, Ngo B H, et al. CLEAR: Cross-Transformers With Pre-Trained Language Model for Person Attribute Recognition and Retrieval[J]. Pattern Recognition, 2025, 164: 111486. <https://doi.org/10.1016/j.patcog.2025.111486>
- [16] Ranjit Kumar Mishra, Arijit Mondal, Jimson Mathew.Nystromformer based cross-modality transformer for visible-infrared person re-identification.Scientific Reports, 2025, Vol. 15(1).<https://doi.org/10.1038/s41598-025-01226-5>.

- [17] Rangwan Kasantikul, Worapan Kusakunniran, Qiang Wu, Zhiyong Wang. Channel-shuffled transformers for cross-modality person re-identification in video General information. 2025, Scientific Reports, <https://doi.org/10.1038/s41598-025-00063-w>.
- [18] Zhang X, Dong K, Cheng D, et al. STWANet: Spatio-Temporal Wavelet Attention Aggregation Network for Remote Sensing Change Detection[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025. <https://doi.org/10.1109/JSTARS.2025.3551093>
- [19] Zhou Y, Xia H, Yu D, et al. Outlier detection method based on high-density iteration[J]. Information Sciences, 2024, 662: 120286. <https://doi.org/10.1016/j.ins.2024.120286>
- [20] Liu W, Jia X, Zhong X, et al. Dynamic and static mutual fitting for action recognition[J]. Pattern Recognition, 2025, 157: 110948. <https://doi.org/10.1016/j.patcog.2024.110948>
- [21] Su Y, Tan Y, An S, et al. Semantic-driven dual consistency learning for weakly supervised video anomaly detection[J]. Pattern Recognition, 2025, 157: 110898. <https://doi.org/10.1016/j.patcog.2024.110898>
- [22] Xu H, Gao Y, Hui Z, et al. Language knowledge-assisted representation learning for skeleton-based action recognition[J]. IEEE Transactions on Multimedia, 2025. <https://doi.org/10.1109/TMM.2025.3543034>
- [23] Su Y, Xing M, An S, et al. Vdarn: Video disentangling attentive relation network for few-shot and zero-shot action recognition[J]. Ad Hoc Networks, 2021, 113: 102380. <https://doi.org/10.1016/j.adhoc.2020.102380>
- [24] He S, Ding L, Xiong Z, et al. A distinctive Eocene Asian monsoon and modern biodiversity resulted from the rise of eastern Tibet[J]. Science Bulletin, 2022, 67(21): 2245-2258. <https://doi.org/10.1016/j.scib.2022.10.006>
- [25] Su Y, Chen D, Xing M, et al. Coming out of the dark: human pose estimation in low-light conditions[C]//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. 2025: 1882-1890. <https://doi.org/10.24963/ijcai.2025/210>
- [26] Sabrina A, Sarra B, and Khawla M. Visual-HAR: A Real-Time Human Activity Recognition Model Based on Visual Sensors. 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Male, Maldives. 2024, 1-6.
- [27] Zhang D, et al. Multi-view and multi-scale behavior recognition algorithm based on attention mechanism. Frontiers in Neurobotics. 2023, 17: 1276208. <https://doi.org/10.3389/fnbot.2023.1276208>
- [28] Yu Xueyang, Li Shuqiu, Ning Jiangbo, et al. A method for underwater acoustic target recognition based on the VGGish neural network [J]. Acoustic Technology. 2024, 43(01):30-37. (in Chinese).
- [29] Luo R.K. et al. A YOLO-I3D: Optimizing Inflated 3D Models for Real-Time Human Activity Recognition. JOURNAL OF IMAGING. 2024, <https://doi.org/10.3390/jimaging10110269>
- [30] Serrano S., Scarpa M.L. et al. VGGISH For Music/Speech Classification in Radio Broadcasting. PROCEEDINGS OF THE 38TH ECMS INTERNATIONAL CONFERENCE ON MODELLING AND SIMULATION ECMS 2024, 550-557
- [31] Li Hang, Chen Zhigang, Wang Yijie, et al. Research on Human Posture Abnormality Detection Based on a Spatiotemporal Graph Attention State Space Model [J]. Computer Engineering and Science. 2025, 47(10):1830-1840. (in Chinese).
- [32] Saghir Alfasy, Jian Lu et al. Auxiliary audio-textual modalities for better action recognition on vision-specific annotated videos. Pattern Recognition, <https://doi.org/10.1016/j.patcog.2024.110808>