2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

ASTPGFN: A Parallel Gating Fusion Framework with Adaptive Spatio-Temporal Modeling for Human Activity Recognition

Yan Mao[1], Guoyin Zhang[1] and Cuicui Ye[2]

[1] College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2] School of Electronic Information, Huzhou College, Huzhou, China
zhangguoyin@hrbeu.edu.cn

Abstract. Effective spatiotemporal modeling is crucial for Human Activity Recognition (HAR) using wearable sensors. This paper proposes a novel HAR method integrating a feature enhancement layer, a spatiotemporal gated fusion module, and a fine-grained spatiotemporal segmentation attention module. The feature enhancement layer transforms input data to improve its representational capacity. The spatiotemporal gated fusion module extracts global spatiotemporal features using a Transformer-based temporal encoder and a residual Graph Convolutional Network (GCN)-based spatial encoder, with an adaptive gating mechanism for feature fusion. The fine-grained segmentation attention module further refines local spatial and temporal features to enhance feature interaction. The fully integrated features are then classified using a fully connected layer. Experimental results on multiple public datasets demonstrate that the proposed method outperforms conventional approaches in terms of recognition accuracy, robustness, and generalization. This model provides an efficient and adaptive solution for HAR using wearable sensors.

Keywords: Wearable sensors, Human Activity Recognition, Spatiotemporal modeling, Graph convolutional network, Transformer.

1 Introduction

Human Activity Recognition (HAR) is a key research area in ubiquitous computing, playing a critical role in applications such as smart homes, health monitoring, sports training, rehabilitation management, and human–computer interaction [1]. With the increasing prevalence of wearable devices such as smartwatches, smartphones, and wristbands, vast amounts of data are continuously generated by embedded multi-channel inertial sensors, including accelerometers, gyroscopes, and magnetometers. These sensor data provide a valuable basis for real-time monitoring and accurate analysis of human motion. Compared to vision-based activity recognition methods, inertial sensors offer notable advantages, including low power consumption, ease of deployment, and strong stability, while also reducing privacy concerns associated with

image capture [2,3]. Therefore, inertial sensor-based HAR has emerged as a major focus in recent research.

Existing Human Activity Recognition (HAR) methods predominantly utilize deep learning models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), Convolutional Neural Networks (CNNs), and Transformers. These approaches have achieved notable progress in temporal sequence modeling. However, most primarily focus on temporal dependencies, often overlooking the spatial correlations inherent in multi-channel inertial sensor data. In human movement, inertial sensor channels exhibit not only temporal dependencies but also potential spatial interactions. Effectively capturing this spatial dependency information is crucial for enhancing recognition accuracy and improving model generalization [4]. Some research have explored the use of graph structures to model cross-channel and cross-sensor dependencies. For instance, Mao et al. [25] constructed adjacency matrices based on expert priors to organize sensor data into multi-level graph structures, and employed Graph Attention Networks (GAT) to extract spatio-temporal features. While these approaches enhance spatial modeling to some extent, they still face two main limitations: (1) predefined graph structures lack adaptability to dynamic changes in activity patterns, limiting the model's ability to capture individual behavioral differences; and (2) such structures cannot dynamically adjust the interaction weights between channels based on the data distribution, resulting in suboptimal spatial feature extraction. Therefore, dynamically modeling inter-channel relationships and integrating them with temporal features remains a core challenge in HAR.

To address the limitations of predefined graph structures and the separate modeling of spatial and temporal features, this paper proposes a human activity recognition method based on adaptive graph learning and parallel spatio-temporal feature extraction. The proposed model consists of three core components: a feature transformation layer, a Spatio-Temporal Gated Fusion Module (STGFM), and a Fine-Grained Spatio-Temporal Segmentation Attention Module (STSAM).

The feature transformation layer enhances the representational capacity of raw sensor data. STGFM integrates a Transformer-based temporal encoder and a residual GCN-based spatial encoder, with a gating mechanism to adaptively fuse temporal and spatial features. The spatial encoder dynamically constructs behavior graphs using learnable embedding vectors, enabling adaptive modeling of inter-channel dependencies. STSAM further captures local spatio-temporal dependencies through segmented attention, enhancing fine-grained feature interactions. Finally, the fused spatio-temporal features are passed through a fully connected layer to achieve accurate human activity classification.

The main contributions of this chapter are as follows:

- A global graph modeling method based on adaptive graph learning is proposed, which dynamically updates the adjacency matrix to model spatial dependencies among multi-channel sensors, effectively addressing the limitations of fixed graph structures.
- A spatio-temporal gated fusion mechanism is designed, integrating a Transformer-based temporal encoder and a GCN-based spatial encoder, enabling

the adaptive fusion of temporal and spatial features and enhancing global representation capability.

- A fine-grained spatio-temporal segmentation attention mechanism is introduced to strengthen local spatio-temporal interactions and improve classification accuracy in activity recognition.

Extensive experiments on four public datasets demonstrate that the proposed method outperforms existing prior-graph-based HAR approaches in terms of accuracy, generalization, and robustness.

The remainder of this paper is organized as follows. Section 2 reviews related studies, focusing on the development of HAR techniques, the application of GNNs in sensor data modeling, and recent advances in automated graph learning. Section 3 details the proposed method. Section 4 presents the experimental setup, including dataset descriptions, evaluation metrics, and result analysis. Section 5 concludes the paper and discusses future research directions.

2 Relate Works

Human Activity Recognition (HAR) involves modeling spatio-temporal dependencies in multi-channel sensor data. A variety of deep learning models have been proposed to address this challenge. CNN1D models are widely used for local temporal feature extraction due to their computational efficiency, though they are limited in capturing long-range dependencies [5-9]. In contrast, CNN 2D models transform sensor signals into image-like representations (e.g., spectrograms or activity images) to enhance spatial modeling [10-13], albeit at higher computational cost. LSTM and GRU models remain popular for temporal modeling. LSTM-based models are capable of learning long-term dependencies and benefit from attention mechanisms [14-18], but are less efficient in real-time applications. GRU offers better efficiency and has been enhanced with dual attention mechanisms and graph integration to improve performance [19-23], though it still lacks explicit spatial modeling. Transformer-based models have shown strong performance in HAR due to global self-attention[24,25], especially when combined with spatial structures like GAT. However, their high computational complexity and large data requirements hinder real-world deployment. Graph Neural Networks (GNNs) have emerged as powerful tools for modeling non-Euclidean spatial structures. Hybrid CNN-GCN models [26] and multi-graph convolutional networks [27] enhance spatial feature learning. However, these approaches often rely on manually defined graphs, limiting their generalizability. Recent work explores automatic graph construction techniques that learn data-driven spatial dependencies [28]. These methods reduce dependence on expert knowledge and improve adaptability. Despite the progress made, most existing approaches either neglect the complex spatial interactions among sensor channels or depend on static, manually defined graph structures that fail to generalize across different activities or sensor configurations. Furthermore, many models treat spatial and temporal dependencies separately, missing the opportunity to jointly model these two aspects in an integrated and adaptive manner. To overcome these limitations, we propose a Adaptive Spatio-Temporal Parallel Gating

Fusion Network (ASTPGFN), which dynamically learns multiple types of graph structures and unifies spatio-temporal attention mechanisms for more expressive and flexible representation learning in HAR tasks.

3 The Proposed Model

This paper proposes a human action recognition method ASTPGFN based on adaptive graph learning and parallel spatiotemporal feature extraction, which aims to efficiently model the spatiotemporal dependencies of multi-channel sensor data. Figure 1 describes the overall framework of the model and details each component that makes up the architecture.

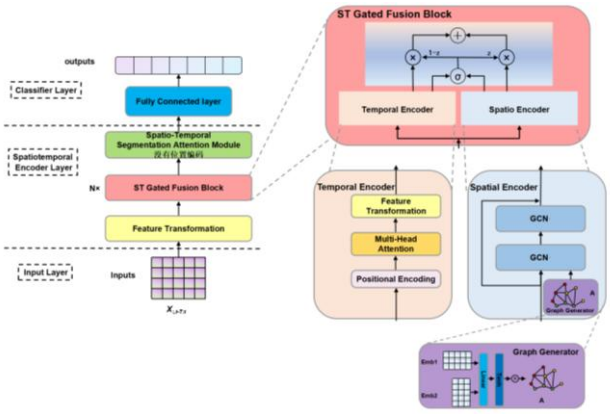

Fig. 1. Overall frame structure.

Aiming at the problem of expert experience required to build spatial structure based on prior knowledge and insufficient description of fine-grained instantaneous changes in behavior, a wearable human behavior recognition spatiotemporal modeling method combining global adaptive behavior graph and local dynamic graph is proposed. This method first automatically builds a spatial behavior graph structure based on the input multi-channel sensor data through an adaptive graph learning module, dynamically captures the spatial dependency between sensor channels, avoids strong dependence on expert domain knowledge, and improves the adaptability and generalization ability of the model. Subsequently, a graph convolutional network (GCN) and an improved lightweight Transformer structure are used to extract spatial and temporal features, respectively, and position encoding is used to retain the time order and original feature expression. Through the spatial temporal gated fusion module (STGFM), dynamic fusion between spatial and temporal features is achieved, and the model's ability to express global spatiotemporal dependencies is enhanced. Furthermore, considering that the local dynamic features of short-term behaviors may be ignored in the global modeling process, this paper divides the time series into multiple time slices and introduces a local channel attention module (LCAM) in each time slice for independent modeling to enhance the model's ability to capture local

behavior changes. At the same time, the channel self-attention mechanism is introduced to dynamically adjust the fusion weights of different sensor channel features, thereby improving the discrimination of cross-channel features. Finally, the fused global and local spatiotemporal feature vector is input into the classifier to complete behavior recognition.

3.1 Adaptive Graph Construction Module

To enhance spatial modeling, ASTPGFN incorporates an adaptive graph generation module [29], which dynamically learns adjacency matrices from input distributions, replacing static, prior-based graph construction and improving adaptability across scenarios.

Specifically, for a sensor network with N nodes, denoted as $V = (v_1, v_2, \cdots, v_N)$,we introduce two trainable node embedding matrices $\mathrm{E}_1, \mathrm{E}_2 \in \mathbb{R}^{\mathrm{N}\times\mathrm{d}}$, where d is the embedding dimension. These embeddings are first transformed via linear projections, and then used to compute an asymmetric interaction matrix, from which the initial adjacency matrix $\mathrm{A} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ is obtained:

$$\mathbf{A} = Softmax(ReLU(E_1 E_2^T)) \tag{1}$$

This formulation captures non-symmetric node relationships and allows the graph structure to be learned end-to-end during training. The use of dual embeddings and nonlinear transformations enables the model to adaptively adjust connection strengths between nodes based on their latent representations, thereby constructing a behavior-aware adjacency graph.

To ensure numerical stability and consistent feature propagation in graph convolution operations, the learned adjacency matrix is further normalized using the symmetric normalization strategy:

$$\widetilde{\mathbf{A}} = \mathrm{D}^{-1/2}\mathrm{A}\mathrm{D}^{-1/2} \tag{2}$$

Where $\mathrm{D} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ is the diagonal degree matrix of A, with $\mathrm{D}_{\mathrm{ii}} = \sum_{\mathrm{j}} \mathrm{A}_{\mathrm{ij}}$. The normalized adjacency matrix $\widetilde{\mathbf{A}}$ is subsequently used in the spatial encoder (graph convolution) to guide the aggregation of node features across the spatial dimension.

3.2 Spatio-Temporal Gated Fusion Module

To capture spatio-temporal dependencies in multi-channel sensor data, ASTPGFN adopts a parallel structure with stackable Spatio-Temporal Gated Fusion Modules (STGFM)[30], each integrating a spatial encoder, temporal encoder, and gated fusion to jointly learn spatial and temporal features.

At each step t, the spatial encoder takes the node features $\mathrm{X}_\mathrm{t} \in \mathbb{R}^{\mathrm{B}\times\mathrm{N}\times\mathrm{D}}$ and applies a two-layer Graph Convolutional Network (GCN) guided by a normalized adjacency matrix $\widetilde{\mathrm{A}}$. To enhance the model's representational capacity, the node features $\mathrm{X}_\mathrm{t} \in \mathbb{R}^{\mathrm{B}\times\mathrm{N}\times\mathrm{D}}$ are projected to a higher-dimensional space through a feature transformation layer using a linear transformation.The computation is as follows:

$$H_t^{(s)} = \sigma\left(\tilde{A}X_t W_1\right) + \tilde{A}\left(\sigma\left(\tilde{A}X_t W_1\right)\right)W_2 \tag{3}$$

where $W_1, W_2 \in \mathbb{R}^{D\times D}$ are trainable weights and $\sigma(\cdot)$ denotes a nonlinear activation function.

For temporal modeling, the temporal encoder takes node sequences as input and employs a Multi-head Self-Attention mechanism to capture global dependencies across time steps. For each attention head k, the computation is defined as follows:

$$\mathrm{Q}^{(\mathrm{k})} = \mathrm{XW}_{\mathrm{q}}^{(\mathrm{k})}, \mathrm{K}^{(\mathrm{k})} = \mathrm{XW}_{\mathrm{k}}^{(\mathrm{k})}, \mathrm{V}^{(\mathrm{k})} = \mathrm{XW}_{\mathrm{v}}^{(\mathrm{k})}$$

$$\mathrm{A}^{(\mathrm{k})} = \mathrm{softmax}\left(\frac{\left(\mathrm{Q}^{(\mathrm{k})}\right)^{\mathrm{T}}\left(\mathrm{K}^{(\mathrm{k})}\right)}{\sqrt{\mathrm{d}}}\right), \mathrm{H}^{(\mathrm{k})} = \mathrm{A}^{(\mathrm{k})}\mathrm{V}^{(\mathrm{k})} \quad (4)$$

The outputs of all K heads are concatenated and linearly transformed to yield the final temporal feature $\mathrm{H}_{\mathrm{t}}^{(\mathrm{t})} \in \mathbb{R}^{\mathrm{B}\times\mathrm{T}\times\mathrm{N}\times\mathrm{D}}$.

To enable effective fusion of spatial and temporal features, STGFM incorporates a gated fusion mechanism. Specifically, a gating vector Z is generated using the Sigmoid activation function to adaptively weight and combine the spatial and temporal features:

$$Z = \sigma\left(H_t^{(s)}W_s + H_t^{(t)}W_t\right) \quad (5)$$

$$H = Z \odot H_t^{(s)} + (1 - Z) \odot H_t^{(t)} \quad (6)$$

where $\odot$ denotes element-wise multiplication, and $W_s, W_t \in \mathbb{R}^{D\times D}$ are trainable parameters.The fused result H serves as the output of the current layer and the input to the next STGFM layer.

This modular design supports multi-layer stacking, enabling progressive abstraction of higher-order spatio-temporal representations with enhanced structural expressiveness and dynamic modeling capability.

3.3 Local Channel Attention Module

After multi-layer spatio-temporal modeling, ASTPGFN introduces a Local Channel Attention Module (LCAM) to enhance local spatio-temporal feature learning. Based on multi-head attention, LCAM independently models temporal dynamics for each sensor node and aggregates outputs into a unified representation.

Let the spatio-temporal feature representation encoded by the STGFM layer be denoted as $\mathrm{H}^{\mathrm{L}} \in \mathbb{R}^{\mathrm{B}\times P\times\mathrm{N}\times F}$, where B is the batch size, P is the number of time slices, N is the number of sensor nodes, and F is the feature dimension. Assume there are h attention heads, each with a head dimension of d, such that $F = h \cdot d$. To enable feature correlation across the temporal dimension for each node, the input feature tensor is reshaped as $H_{perm} \in \mathbb{R}^{B\times F\times N\times P}$.

This reshaping enables each sensor node to perform interaction modeling across all time slices.

Based on this structure, three sets of fully connected layers with shared parameters are applied to generate the Query, Key, and Value representations, respectively:

$$Q = FC_q\left(H_{perm}\right), \quad K = FC_k\left(H_{perm}\right), \quad V = FC_v\left(H_{perm}\right) \quad (7)$$

Each of these representations has the shape $\mathbb{R}^{\mathrm{B}\times\mathrm{F}\times\mathrm{N}\times\mathrm{P}}$,enabling each attention head to learn temporal response features across different time steps for each sensor node.

The feature dimension F is split into K attention heads, each with a dimension of d, and the resulting head-wise representations are concatenated along the batch dimension to form the multi-head representation:

$$Q_h, K_h, V_h \in \mathbb{R}^{KB \times d \times N \times P} \tag{8}$$

To match the matrix multiplication format required for attention computation, the dimensions of Q_h, K_h, V_h are reshaped to $\mathbb{R}^{KB \times N \times P \times d}$, enabling the calculation of scaled attention weights. At each sensor node, the attention mechanism independently models the internal relationships across time slices. The attention weights are computed as:

$$\text{Att} = \text{softmax}\left(\frac{Q_h {K_h}^{\top}}{\sqrt{\text{D}}}\right) \in \mathbb{R}^{KB \times N \times P \times P} \tag{9}$$

The Softmax operation is applied along the temporal dimension, allowing each node to dynamically attend to features from different time steps. The attention weights are then applied to the value vectors to obtain the attention-weighted output:

$$Z_h = Att \cdot V_h \in \mathbb{R}^{KB \times N \times P \times d} \tag{10}$$

The outputs from all attention heads are reshaped back to the original batch size, transposed, and concatenated along the feature dimension to form the fused representation $\text{Z} \in \mathbb{R}^{\text{B} \times P \times \text{N} \times F}$.To aggregate the multi-head attention outputs and enhance non-linear representational capacity, a fully connected layer followed by a non-linear activation is applied:

$$H' = \sigma(ZW + b) \in \mathbb{R}^{B \times P \times N \times F} \tag{11}$$

Where W and b are learnable parameters, and $\sigma(\cdot)$ denotes a non-linear activation function (e.g., ReLU).H' serves as the final output of the LCAM module.

3.4 Classifier

The feature representations generated by the STSAM module are fed into a fully connected layer, followed by global average pooling across temporal and spatial dimensions to obtain a sample-level spatio-temporal representation, which is then used for activity classification via a linear classifier. ASTPGFN is trained using the cross-entropy loss function, optimized through backpropagation in an end-to-end supervised manner. To enhance training stability and convergence, the model incorporates Batch Normalization, weight initialization, and learning rate decay strategies.

4 Experiments and Results

4.1 Dataset

To evaluate the performance of the ASTPGF model, four widely used open-source datasets were employed: UCI-HAR, MotionSense, Shoaib, and WISDM. Table 1 summarizes the detailed information of each dataset. The activities included in the datasets are: walking (wlk), upstairs (ups), downstairs (dws), sitting (sit), standing (std), lying (lay), jogging (jog), and biking (bik).

4.2 Experiment details

The experiments were conducted on a server equipped with an Intel Core i9-9900K processor running at 3.6 GHz, 32 GB of RAM, and the Windows 10 (x64) operating system. The proposed model was implemented and evaluated using the Python programming language. The batch size was set to 64, and the maximum number of

training epochs was 100. The Adam optimizer was employed for training. A learning rate of 1e-3 was used for all datasets except for the Shoaib dataset, which used a learning rate of 1e-4. The weight decay was set to 1e-3. All experiments were performed on an NVIDIA 2080Ti GPU.

Table 1. Statistics of the datasets.

Dataset	Subject	SR	Channels	Sample	Activity	SW	Activities
UCI-HAR	30	50	9	1318272	6	128	wlk,ups,dws,sit,std,lay
Motion	24	50	12	1412865	6	120	dws,jog,sit,std,ups,wlk
Shoaib	10	50	9	63000	7	120	wlk,sit,std,job,bik,ups,dws
WISDM	36	20	3	1098208	6	120	wlk,jog,sit,std,ups,dws

4.3 Evaluation Metrics

$$\textbf{Acuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

Where TP and TN are true positive and true negative rates, FP and FN are false positive and false negative rates.

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{3}$$

We use macro average F1-score as metric to compare the performance of the proposed method with other methods. In this regard, we calculate F1-score for each class according to equation (4) as follows:

$$\textbf{MacroF1} - \textbf{Score} = \frac{2}{|C|} \times \sum_{i=1}^{C} \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{15}$$

Where, $|C|$ in equation(15) indicates number of classes and F1-score for each class is given the same weight irrespective of their number of instances and $i = 1,2,\cdots C$ is considered as the set of classes considered for experiment.

$$\textbf{MicroF1} - \textbf{Score} = 2\sum_{i=1}^{C} \frac{N_c}{N_{total}} \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{16}$$

Micro F1-score is different from macro F1-score, weighted by the number of samples in different class.

4.4 Compare with other experiments

To validate the effectiveness of the proposed method, this section conducts a comparative evaluation against established baseline models for human activity recognition.

- Transformer[24]: Adapted from the original Transformer[31], this model modifies input/output formats for activity recognition and leverages parallel attention for more efficient temporal modeling than RNNs..
- SelfAttention[32]: This method uses a Transformer-like structure with self-attention and 1×1 convolutions for multi-channel fusion, integrating multi-modal data and global temporal attention to extract key spatio-temporal features.
- GraphConvLSTM[33]: The model combines GCN and TCN to extract spatial-temporal features from skeletal graphs, followed by LSTM for global modeling and classification.
- DDNN[32]: Integrates LSTM and CNN with feature extraction from temporal, spatial, and statistical domains, followed by feature concatenation for activity recognition.
- ConvLSTM[35]: Stacks convolutional layers to extract spatial features and LSTM layers for temporal modeling in activity recognition.
- STGTN(FC)[25]: Utilizes a fully connected channel-level adjacency matrix as input.
- STGTN(VLG)[25]: Uses locally fully connected adjacency matrices based on sensor nodes, with additional virtual local and global nodes.

To evaluate component effectiveness, the proposed ASTPGFN model is tested under different graph configurations:

- ASTPGFN (Directed): Uses directed graphs based on sensor nodes as input.
- ASTPGFN (Undirected): Uses undirected graphs based on sensor nodes as input.

Table 2 shows that ASTPGFN, in both its directed and undirected forms, consistently outperforms existing baseline methods across the UCI-HAR, Motion, Shoaib, and WISDM datasets, demonstrating strong generalization under diverse sensor configurations and sampling conditions. The directed variant achieves better performance, likely due to its ability to capture asymmetric information flow between sensor nodes, whereas the undirected structure loses directional cues, limiting its capacity to model complex temporal-spatial relationships. Directionality is particularly important in activity recognition, where signal propagation exhibits inherent temporal order. ASTPGFN integrates adaptive graph structures and directed connections to dynamically model sensor dependencies, enabling effective extraction of key patterns and suppression of irrelevant features, especially in high-resolution datasets like UCI-HAR and Motion. Compared to Transformer-based models and fixed-window approaches like ConvLSTM and DDNN, ASTPGFN optimizes spatio-temporal information flow via adaptive topologies, showing notable advantages on datasets such as Shoaib, where class balance coexists with limited subject diversity. Unlike GNN models with predefined adjacency (e.g., GraphConvLSTM, STGTN), ASTPGFN dynamically adjusts inter-sensor connectivity, improving robustness in complex and heterogeneous sensing scenarios. Its performance on the low-sampling WISDM dataset further highlights its capability to extract meaningful features under limited temporal resolution. Overall, ASTPGFN offers a robust and generalizable framework for human activity recognition across diverse and challenging data environments.

Figure 2 presents the confusion matrices of the ASTPGFN model using directed weighted graphs across four different datasets. The results show that the model achieves

Table 2. Comparison of baseline results(mif/maf).

Methods	Motion	UCI-HAR	Shoaib	Wisdm
ASTPGFN（Directed）	**0.9688/0.9604**	**0.9206/0.9204**	**0.9542/0.9532**	**0.8668/0.8249**
ASTPGFN(Undirected)	0.9640/0.9535	0.9175/0.9176	0.9476/0.9471	0.8564/0.8117
STGTN (FC)[25]	0.9629/0.9510	0.9138/0.9131	0.9429/0.9397	0.8627/0.8248
STGTN (VLG)[25]	0.9658/0.9554	0.9165/0.9161	0.9524/0.9523	0.8565/0.8110
Transformer[24]	0.9294/0.9044	0.9104/0.9088	0.9067/0.8947	0.8133/0.7615
GraphConvLSTM[33]	0.9435/0.9284	0.9091/0.9092	0.9257/0.9246	0.8179/0.7644
DDNN[34]	0.9554/0.9419	0.9057/0.9051	0.8952/0.8777	0.8526/0.8103
ConvLSTM[35]	0.9539/0.9433	0.9053/0.9048	0.8895/0.8860	0.8549/0.8071
SelfAttention[32]	0.9272/0.9086	0.8949/0.8935	0.7010/0.6838	0.7361/0.6778

high classification accuracy for most activity categories, with particularly stable performance in recognizing dynamic activities such as “jogging,” where the accuracy approaches 1.0. Static activities like “sitting” and “standing” are also classified with high accuracy, though some confusion between the two is observed in certain datasets. Misclassification is more prominent between “upstairs” and “downstairs,” indicating significant similarity in their inertial sensor patterns. Additionally, some “walking” samples are misclassified as stair-related activities, suggesting overlapping acceleration trends between walking and stair movements. Overall, the ASTPGFN model outperforms baseline methods across multiple datasets, demonstrating strong classification capability under various human activity patterns.

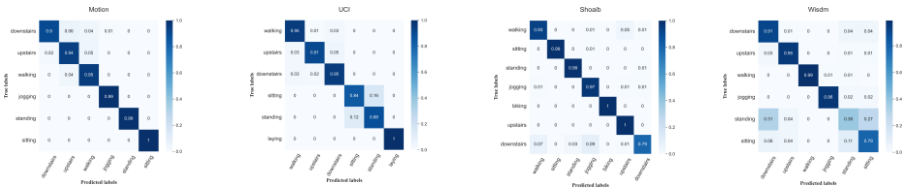

Fig. 2. Confusion matrix.

4.5 Ablation experiment

Table 3 reports the classification performance of individual modules to evaluate their standalone contributions and the overall model effectiveness. A “leave-one-module-in”

ablation strategy is adopted, where only one module is retained at a time. The full model, ASTPGFN, serves as the performance upper bound.

As shown in Table 3, both the spatial and temporal modules consistently contribute to stable performance across datasets, highlighting their essential roles in human activity recognition. Their relative importance varies by dataset: the temporal module performs better on the Shoaib dataset due to more distinctive temporal patterns, while the spatial module is generally more effective elsewhere. In contrast, models retaining only the fusion module exhibit significantly lower performance, indicating that without explicit spatial or temporal modeling, feature fusion lacks discriminability and may introduce noise. The LCAM module, which constructs fully connected intra-slice graphs via a Transformer-based approach, achieves moderate performance without dynamic graph construction, suggesting structural consistency but limited capacity for capturing spatio-temporal dependencies. The complete model consistently outperforms all variants, demonstrating the complementarity of its components and the effectiveness of joint modeling of spatial, temporal, and global structures. These findings validate the model architecture and highlight the necessity of multi-dimensional feature integration for optimal recognition performance.

Table 3 Comparison results of model ablation methods(mif/maf).

Graph type	Models	Motion	UCI	Shoaib	Wisdm
Directed	Only Spatial	0.9450/0.9309	0.8870/0.8856	0.9028/0.9040	0.8445/0.8065
	Only Temporal	0.9112/0.8852	0.8639/0.8610	0.9419/0.9412	0.8252/0.7667
	Only Fusion	0.3154/0.2716	0.4086/0.3575	0.5472/0.4980	0.6403/0.4708
	Only LCAM	0.9361//0.9217	0.8731/0.8726	0.9113/0.9118	0.8499/0.8078
	ASTPGFN	**0.9688/0.9604**	**0.9206/0.9204**	**0.9542/0.9532**	**0.8668/0.8249**
Undirected	Only Spatial	0.9525/0.9420	0.8745/0.8735	0.8885/0.8874	0.8557/0.8001
	Only Temporal	0.9261/0.9053	0.8582/0.8553	0.9457/0.9457	0.8082/0.7539
	Only Fusion	0.3830/0.3799	0.4276/0.3756	0.5434/0.4822	0.6731/0.5049
	Only LCAM	0.9495/0.9386	0.8677/0.8665	0.8847/0.8807	0.8425/0.7649
	ASTPGFN	**0.9640/0.9535**	**0.9175/0.9176**	**0.9476/0.9471**	**0.8564/0.8117**

5 Conclusion

In this paper, we propose a spatio-temporal multi-attention recognition model based on graph neural networks to address the limitations in spatial dependency modeling and temporal feature extraction in multi-channel sensor-based human activity recognition. An adaptive graph generation mechanism is introduced to dynamically construct four types of adjacency matrices, capturing complex spatial relationships among sensor channels. Multi-head temporal attention and gated fusion modules enhance temporal dynamics modeling and feature integration. Additionally, local and channel-wise attention mechanisms improve the representation of local and cross-channel features. Experiments on public datasets demonstrate that the proposed method achieves superior accuracy, robustness, and adaptability, particularly in modeling directional dependencies with asymmetric weighted graphs.

References

1. Bulling A, Blanke U, Schiele B. A tutorial on human activity recognition using body-worn inertial sensors[J]. ACM Computing Surveys, 2014, 46(3): 1–33.
2. ZHENG Y-D, LIU Z, LU T, et al. Dynamic Sampling Networks for Efficient Action Recognition in Videos [J]. IEEE Transactions on Image Processing, 2020, 29: 7970-83.
3. TANG Y, TIAN Y, LU J, et al. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition [Z]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 5323-32.10.1109/cvpr.2018.00558
4. Zhan J. Single-Device Human Activity Recognition Based on Spatiotemporal Feature Learning Networks[J]. Transactions on Computational and Scientific Methods, 2025, 5(3).
5. Ignatov A. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. Appl Soft Comput, 2018, 62: 915～922
6. Ronao C A, Cho S B. Deep convolutional neural networks for human activity recognition with smartphone sensors[C]//Neural Information Processing: 22nd International Conference, ICONIP 2015, November 9-12, 2015, Proceedings, Part IV 22. Springer International Publishing, 2015: 46-53.
7. Yang J, Nguyen M N, San P P, et al. Deep convolutional neural networks on multichannel time series for human activity recognition[C]//Ijcai. 2015, 15: 3995-4001.
8. Ha S, Choi S. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors[C]//2016 international joint conference on neural networks (IJCNN). IEEE, 2016: 381-388.
9. Ronao C A, Cho S B. Human activity recognition with smartphone sensors using deep learning neural networks[J]. Expert systems with applications, 2016, 59: 235-244.
10. Jiang W, Yin Z. Human activity recognition using wearable sensors by deep convolutional neural networks[C]//Proceedings of the 23rd ACM international conference on Multimedia. 2015: 1307-1310.
11. Ha S, Yun J M, Choi S. Multi-modal convolutional neural networks for activity recognition[C]//2015 IEEE International conference on systems, man, and cybernetics. IEEE, 2015: 3017-3022.
12. Ito C, Cao X, Shuzo M, et al. Application of CNN for human activity recognition with FFT spectrogram of acceleration and gyro sensors[C]//Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers. 2018: 1503-1510.
13. Xi R, Hou M, Fu M, et al. Deep dilated convolution on multimodality time series for human activity recognition[C]//2018 international joint conference on neural networks (IJCNN). IEEE, 2018: 1-8.
14. Zeng M, Gao H, Yu T, et al Understanding and improving recurrent networks for human activity recognition by continuous attention. 见: Proceedings of the 2018 ACM International Symposium on Wearable Computers. New York, NY, USA: Association for Computing Machinery, 2018. 56～63
15. Yao S, Hu S, Zhao Y, et al. Deepsense: A unified deep learning framework for time-series mobile sensing data processing[C]//Proceedings of the 26th international conference on world wide web. 2017: 351-360.
16. Zhao Y, Yang R, Chevalier G, et al. Deep residual bidir-LSTM for human activity recognition using wearable sensors[J]. Mathematical problems in engineering, 2018, 2018(1): 7316954.
17. Gumaei A, Hassan M M, Alelaiwi A, et al. A hybrid deep learning model for human activity recognition using multimodal body sensing data[J]. IEEE Access, 2019, 7: 99152-99160.

18. Xu C, Chai D, He J, et al. InnoHAR: A deep neural network for complex human activity recognition[J]. Ieee Access, 2019, 7: 9893-9902.
19. Dua N, Singh S N, Semwal V B. Multi-input CNN-GRU based human activity recognition using wearable sensors[J]. Computing, 2021, 103(7): 1461-1478.
20. Lu L, Zhang C, Cao K, et al. A multichannel CNN-GRU model for human activity recognition[J]. IEEE Access, 2022, 10: 66797-66810.
21. Pan J, Hu Z, Yin S, et al. GRU with dual attentions for sensor-based human activity recognition[J]. Electronics, 2022, 11(11): 1797.
22. Dua N, Singh S N, Semwal V B, et al. Inception inspired CNN-GRU hybrid network for human activity recognition[J]. Multimedia Tools and Applications, 2023, 82(4): 5369-5403.
23. Wu J, Huang J, Wu X, et al. A novel graph-based hybrid deep learning of cumulative GRU and deeper GCN for recognition of abnormal gait patterns using wearable sensors[J]. Expert Systems with Applications, 2023, 233: 120968.
24. Dirgová Luptáková I, Kubovčík M, Pospíchal J. Wearable sensor-based human activity recognition with transformer model[J]. Sensors, 2022, 22(5): 1911.
25. Mao Y, Zhang G, Ye C. A Spatio-temporal Graph Transformer driven model for recognizing fine-grained data human activity[J]. Alexandria Engineering Journal, 2024, 104: 31-45.
26. Wu J, Liu Q. A novel spatio-temporal network of multi-channel CNN and GCN for human activity recognition based on ban[J]. Neural Processing Letters, 2023, 55(8): 11489-11507.
27. Chen L, Luo Y, Peng L, et al. A multi-graph convolutional network based wearable human activity recognition method using multi-sensors[J]. Applied Intelligence, 2023, 53(23): 28169-28185.
28. Chen Y, Qin Y, Li K, et al. Adaptive spatial-temporal graph convolution networks for collaborative local-global learning in traffic prediction[J]. IEEE Transactions on Vehicular Technology, 2023, 72(10): 12653-12663
29. Wu Z, Pan S, Long G, et al. Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. arXiv, 2020.
30. Zheng C, Fan X, Wang C, et al. GMAN: A Graph Multi-Attention Network for Traffic Prediction[C]//arXiv, 2019, 34.
31. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000–6010.
32. Mahmud S, Tonmoy MTH, Bhaumik KK. Human Activity Recognition from Wearable Sensor Data Using Self-Attention[C]//ECAI 2020. IOS Press, 2020: 1332–1339.
33. Han J, He Y, Liu J, et al. GraphConvLSTM: Spatiotemporal Learning for Activity Recognition with Wearable Sensors[C]//2019 IEEE Global Communications Conference (GLOBECOM). Waikoloa, HI, USA: IEEE Press, 2019: 1–6.
34. Qian H, Pan SJ, Da B, et al. A Novel Distribution-Embedded Neural Network for Sensor-Based Activity Recognition[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China: International Joint Conferences on Artificial Intelligence Organization, 2019: 5614–5620.
35. Ordóñez F, Roggen D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition[J]. Sensors, 2016, 16(1): 115.