GC-Trans-UAVNet: A Hybrid GCN-Transformer Framework for Skeleton-Based Action Recognition

Zijian Li*, **Yonglin Liu, Xinye Li**
Harbin Institute of Technology, Weihai
{2022212023,2022212026,2022211871}@stu.hit.edu.cn

Abstract

We present **GC-Trans-UAVNet**, a hybrid framework combining Graph Convolutional Networks (GCN) and Transformers for skeleton-based action recognition. By leveraging GCNs to capture spatial features and Transformers to model temporal dynamics, **GC-Trans-UAVNet** effectively learns both local and global action patterns. Our model has demonstrated strong performance on both the preliminary and final competition datasets. As of the time of writing, our team "T-Rex harvests crops" ranks among the top on the final leaderboard, showcasing **GC-Trans-UAVNet**'s strong performance on the competition's designated tasks.

1 Introduction

Skeleton-based action recognition has gained substantial attention in recent years, especially with the increasing availability of human skeletal data from diverse modalities (Li et al., 2021). Traditional approaches often rely on Graph Convolutional Networks (GCNs) due to their ability to model skeleton data as a spatio-temporal graph, efficiently capturing dependencies between joints over time (Yan et al., 2019; Shi et al., 2019). However, while GCNs excel at learning local spatial features, they often fall short in capturing long-range temporal dependencies, which are critical for recognizing complex, dynamic actions (Liu et al., 2024a; Chen et al., 2021).

In contrast, Transformer architectures have demonstrated superior performance in capturing global temporal patterns, making them particularly suitable for modeling sequential data (Plizzari et al., 2021; Zhang et al., 2021). Transformers are adept at capturing long-term dependencies, which helps address some limitations inherent to GCN-based methods. Nonetheless, Transformers typically struggle with learning localized spatial relationships, especially in the absence of direct, short-term connections between nodes.

To address these challenges, we propose **GC-Trans-UAVNet**, a hybrid framework that combines the strengths of Graph Convolutional Networks (GCNs) and Transformers. As shown in fig. 1, GC-Trans-UAVNet leverages GCNs to capture local spatial dependencies within each frame and Transformers to model long-term temporal dynamics across frames. This selective hybridization effectively integrates both local and global action patterns, enhancing the model's robustness in recognizing complex actions.

Our contributions are threefold:

- We propose a novel dual-branch architecture that selectively integrates GCN and Transformer components, enhancing both spatial and temporal feature learning for skeleton-based action recognition.
- We modify the provided dataset into a three-view architecture to train SKE-MIXF, wherein each view corresponds to a distinct aspect or representation of the original data. The second view may capture the core features essential for the primary task, while the first and third views could encompass auxiliary information or alternative feature representations that support model robustness and generalization.
- We demonstrate the effectiveness of our model on the competition dataset, showing its ability to handle skeletal data captured from aerial views with minimal visual context.
- GC-Trans-UAVNet achieves outstanding performance, ranking among the top on the competition's leaderboard, thereby validating its robustness and accuracy on UAV-based human action recognition tasks.

By integrating the distinct strengths of GCNs and Transformers, GC-Trans-UAVNet marks a

* Corresponding author.

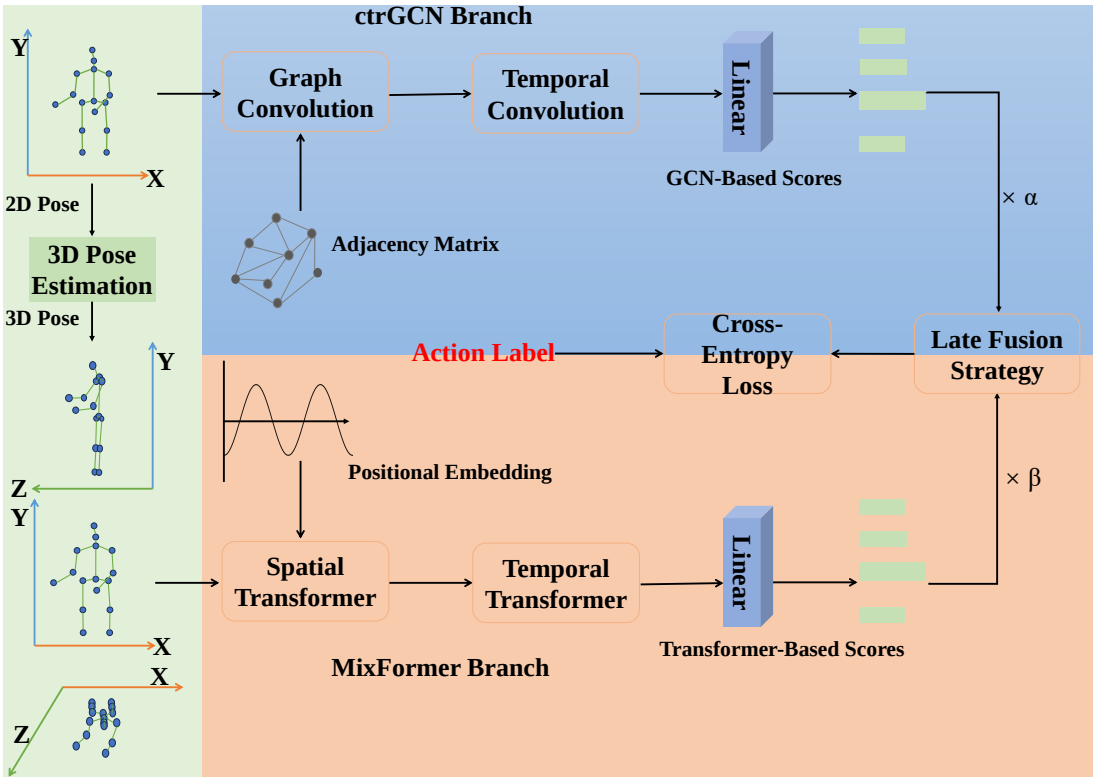


Figure 1: Structure of **GC-Trans-UAVNet**: A two-branch architecture comprising the ctrGCN branch, which leverages GCN and temporal convolutions on 2Dpose data, and the MixFormer branch, which applies spatial-temporal processing with positional embeddings on 3Dpose data. The scores from each branch are fused with distinct weights, and the model is optimized using cross-entropy loss with action labels.

substantial advancement in skeleton-based action recognition, where capturing both spatial and temporal features is essential. The code and detailed training logs are available at [1].

2 Related Work

2.1 GCN-based Methods

As a powerful tool for modeling skeleton data naturally as graphs, GCN is able to effectively capture spatial and temporal dependencies in skeleton data. Yan et al. (Yan et al., 2019) introduced the Spatial-Temporal Graph Convolutional Network (ST-GCN), effectively capturing the spatial dependencies between joints and temporal evolution. Shi et al. (Shi et al., 2019) introduced the Two-Stream Adaptive Graph Convolutional Networks (2s-AGCN), which includes a learnable adjacency matrix and proposes a dual-stream framework for first-order and second-order information fusion, further improving the performance. Cheng et al. (Cheng et al., 2020) proposed the Shift-GCN method, which reduces the computational cost of traditional graph convolution and improves the model performance through a simple and efficient channel-wise shift operation. Liu et al. (Liu et al., 2024a) proposed the Temporal Decoupling Graph Convolutional Network (TD-GCN), which uses temporal-dependent adjacency matrices for temporal-sensitive topology learning from skeleton joints. Chen et al. (Chen et al., 2021) proposed a novel Channel-wise Topology Refinement Graph Convolution (CTR-GCN) which dynamically learns different topologies and effectively aggregates joint features in different channels for skeleton-based action recognition. Cheng et al. (Cheng et al., 2024) proposed a Dense-Sparse Complementary Network (DSCNet), which leverages the complementary information of RGB and skeleton modalities with low computational cost to achieve competitive action recognition performance.

2.2 Transformer-based Methods

The Transformer model is capable of capturing global skeletal action patterns, rather than relying solely on local information. Additionally, it effectively captures dependencies in long sequences, demonstrating significant advantages, especially

[1] https://github.com/ionicbond1234/global_match

when handling large-scale data. Plizzari et al. (Plizzari et al., 2021) applied Transformer to spatial-temporal skeleton-based architectures, through a Spatial Self-Attention (SSA) module and a Temporal Self-Attention (TSA) module. Zhang et al. (Zhang et al., 2021) applied the Transformer model across both spatial and temporal dimensions, proposing a 3D positional encoding address the representation of spatial information among nodes, thereby enabling the Transformer to be applied to graph data. Furthermore, Zhang introduces a spatiotemporal Transformer (ST-TR) to extract spatiotemporal features from skeletal data, facilitating accurate action recognition.

2.3 Hybrid Dual-Branch Methods

Compared to the Graph Convolutional Network (GCN) methods, the Transformer architecture can rapidly capture global topological information and strengthen the association of non-physically connected joints through iterative network updates. However, the Transformer exhibits limitations in distinguishing local features and capturing short-term temporal information. To address these shortcomings, when the Transformer was first applied to skeleton-based action recognition, GCN and Convolutional Neural Networks (CNNs) were incorporated as complementary methods. This integration led to the development of a model termed the hybrid Transformer (Xin et al., 2023), which leverages the strengths of each approach. Liu et al. (Liu et al., 2024b) proposed a novel dual-branch framework called the Hybrid Dual-Branch Network (HDBN). By inputting skeleton data separately into GCN and Transformer backbones and modeling high-level features in parallel, this framework effectively combines the strengths of both architectures. A post-fusion strategy is then applied to further enhance skeleton-based action recognition performance.Differentfrom the method above ,our proposed model improves the dual-branch approach by implementing selective hybridization. Specifically, we input 2D data into the GCN while feeding 3D data into the Transformer, rather than adopting a fully hybrid approach.

3 Skeleton-based Action Recognition

3.1 Preliminary: Task Definition

The objective of this task is to accurately classify human actions based on UAV-captured skeletal data. This data, limited to skeletal modality, provides spatial information on key human joints without detailed visual context, posing unique challenges for model robustness and effective feature extraction.

Mathematically, an action sequence is defined as $S = \{G_t\}_{t=1}^{T}$, where each $G_t = (V_t, E_t)$ represents a spatial graph at time t. Here, $V_t = \{v_i^t \mid i = 1, \ldots, N\}$ denotes the set of N joints (nodes), and $E_t = \{e_{ij}^t \mid i, j = 1, \ldots, N\}$ represents the set of bones (edges) connecting these joints. Each joint v_i^t is specified by coordinates (x_i^t, y_i^t) in 2D Euclidean space.

Previous studies have shown that 3D skeleton data provides more detailed spatial information compared to 2D data (Yan et al., 2019; Shi et al., 2019). Here each joint is represented as (x_i^t, y_i^t, z_i^t) in 3D space.

To capture both spatial dependencies e_{ij}^t (i.e., bones) and the temporal dynamics across frames, we employ a hybrid model integrating Graph Convolutional Networks (GCNs) and Transformer architectures. The GCNs are designed to capture local structural patterns within each frame, while the Transformer layers focus on global temporal dependencies, ensuring comprehensive feature learning for robust action classification.

3.2 Dataset

The dataset used in this study is a subset of the standard benchmark for UAV-based human action recognition, UAV-Human (Li et al., 2021), and contains exclusively skeletal data, structured as follows:

Modalities:

1. **Joint Modality (Primary)**: This modality provides the core spatial coordinates of joints. For each joint v_i in the skeleton sequence S, the coordinates are given as $v_i = (x_i, y_i)$ for 2D data and $v_i = (x_i, y_i, z_i)$ for 3D data.
2. **Bone Modality**: This modality captures pairwise relations between joints, defined as $e_{ij} = (x_i - x_j, y_i - y_j)$ for each pair of connected joints (v_i, v_j) in 2D or $e_{ij} = (x_i - x_j, y_i - y_j, z_i - z_j)$ in 3D.
3. **Motion Modality**: Temporal differences between consecutive frames provide the motion information. The joint motion for each joint v_i across frames t and $t + 1$ is defined as:

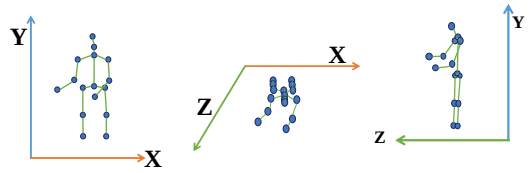
$$m_i^t = v_i^{t+1} - v_i^t = (x_i^{t+1} - x_i^t, y_i^{t+1} - y_i^t).$$

Similarly, the motion of a bone e_{ij} across frames is defined as:

$$m_{ij}^t = e_{ij}^{t+1} - e_{ij}^t.$$

Preprocessing and Modality Generation The dataset preprocessing pipeline leverages provided scripts to generate additional modalities, including bone and motion data, alongside the joint data. This process creates a richer, multi-modal input that captures both spatial and temporal patterns across frames, enhancing the model's ability to recognize complex actions. By combining joint, bone, and motion modalities, we establish a comprehensive foundation for evaluating the effectiveness of GCN and Transformer models in skeleton-based action recognition tasks.

In addition to the core modalities, as shown in fig. 2, we modify the dataset into a three-view architecture for training SKE-MIXF. Each view represents a distinct aspect of the original data: the first and third views offer auxiliary information or alternative feature representations, while the second view focuses on the core features crucial for the primary action recognition task. This multi-view approach promotes model robustness and generalization, making the dataset an ideal platform for testing and refining hybrid architectures that address both global and local feature representation challenges, particularly in UAV-based human action recognition.



(a) First View (b) Second View (c) Third View

Figure 2: Three-view architecture of the dataset for training SKE-MIXF.

3.3 Method

This section details the architecture and components of our proposed GC-Trans-UAVNet model, which integrates innovative data processing techniques, Multi-Source Channel-wise Topology Refinement (MS-CTR-GCN), Three-views-former, and an ensemble strategy to optimally leverage the strengths of multiple data modalities for enhanced skeleton-based action recognition.

Multi-Data Processing Our approach begins with a comprehensive multi-data processing pipeline designed to handle various data modalities. The skeletal action dataset encompasses Joint, Bone, and Motion modalities, each providing distinct spatial and temporal cues essential for action classification:

- **Joint Modality**: Extracts core spatial coordinates of skeleton joints, offering foundational spatial information.
- **Bone Modality**: Captures spatial relationships between joints, providing additional contextual structure and aiding in perceiving complex poses.
- **Motion Modality**: Accounts for temporal dynamics by computing temporal differences, essential for understanding motion across frames.

These modalities are preprocessed to extract meaningful features that are fed into our dual-branch architecture, ensuring comprehensive data utilization.

Multi-Source Channel-wise Topology Refinement GCN (MS-CTR-GCN) The MS-CTR-GCN is a pivotal component of our framework, engineered to adaptively refine the graph topology for skeleton-based representations:

- **Channel-wise Topology Learning**: Dynamically adjusts topological connections within each channel, optimizing the network's ability to capture localized spatial dependencies.
- **Cross-Modality Integration**: Incorporates data from all modalities, allowing for a robust multi-source feature refinement. This integration substantially improves the spatial feature learning by compensating for missing or noisy data in any single modality.

Three-Views-Former The Three-views-former is a Transformer-based module specifically designed to model long-range temporal dependencies through a diversified perspective approach:

- **View 1: Auxiliary Information View**: Enriches the model with additional features that enhance the understanding of complex actions.
- **View 2: Core Feature View**: Captures essential temporal patterns necessary for task execution, focusing on the fundamental motion dynamics.

- **View 3: Alternative Perspective View**: Provides support for model robustness by incorporating diverse feature representations.

This multifaceted perspective ensures holistic temporal feature extraction, improving generalization and accuracy.

Ensemble Strategy To achieve superior action recognition performance, we adopt an ensemble strategy that aggregates predictions from multiple GC-Trans-UAVNet models, each trained on different views or initial conditions:

- **Model Fusion**: Utilizes a weighted voting mechanism where models contribute predictions based on their specialization (e.g., spatial-focused vs. temporal-focused).
- **Robustness Enhancement**: The ensemble approach mitigates individual model biases and errors, promoting accuracy and robustness across varied datasets.

By handling each aspect with a dedicated structural component, GC-Trans-UAVNet consolidates multifarious skeletal data insights, leading to a paradigm shift in how complex actions are recognized from limited visual cues.

4 Experiments

In Table 1, we present the results of our model, **GC-Trans-UAVNet**, compared to several existing methods for skeleton-based action recognition. All experiments were conducted on a setup with 3 NVIDIA-4090 GPUs.

Method	Type	Score↑
MS-CTR-GCN	GCN (joint)	43.70
MS-CTR-GCN	GCN (bone)	42.85
MS-CTR-GCN	GCN (motion)	40.25
MS-CTR-GCN	GCN (longtail)	43.05
TD-GCN	GCN (joint)	42.90
SKE-FORMER	GCN (bone1)	28.00
SKE-FORMER	GCN (bone2)	42.10
SKE-FORMER	GCN (bone3)	33.60
SKE-FORMER	GCN (joint)	41.90
GC-Trans-UAVNet	**Hybrid**	**49.66**

Table 1: Performance of various methods on the given skeleton-based action recognition tasks.

Our model, **GC-Trans-UAVNet**, achieves the highest score of 49.66, surpassing GCN-based models such as **MS-CTR-GCN** and **TD-GCN** (trained on joint features), which reach maximum scores of 43.70 and 42.90, respectively. These results indicate that while GCNs effectively capture local spatial-temporal dependencies, they may lack the ability to model complex, long-range relationships required for detailed action recognition.

The **SKE-FORMER** (transformer-based) displays a range of scores from 28.00 to 42.10, showing sensitivity to different skeleton features. However, transformer models, when combined with GCNs in an ensemble, significantly enhance performance by capturing a broader range of dependencies. This hybrid approach demonstrates that integrating GCNs with transformers, as in **GC-Trans-UAVNet**, yields a robust model capable of comprehensive action recognition across varied feature types.

5 Conclusion

In this work, we introduced GC-Trans-UAVNet, a pioneering hybrid framework that marries the strengths of Graph Convolutional Networks and Transformers to advance skeleton-based action recognition. Through a dual-branch architecture, our model effectively leverages GCNs to discern local spatial dependencies, while Transformers adeptly capture long-range temporal patterns. This integration results in enhanced learning of complex action dynamics, especially in scenarios devoid of rich visual data, like those captured from UAV perspectives.

The three-view architecture used during training, featuring distinct data representations, not only bolsters model robustness but also promotes generalization by encompassing both essential and auxiliary information. This approach ensures that GC-Trans-UAVNet is not just proficient in primary tasks but exhibits marked improvements across varied datasets and scenarios.

Demonstrated by our top-ranked performance in competitive settings, GC-Trans-UAVNet proves its capability in handling sophisticated action recognition tasks. The consistent leaderboard success underscores its potential real-world applicability, particularly in fields reliant on UAV technology for human activity monitoring.

We anticipate that GC-Trans-UAVNet's effective combination of GCNs and Transformers will inspire future work in the domain, driving further innovations in models that require nuanced understanding of spatial and temporal data. Future research may explore broader applications, enhance

data preprocessing methods, or refine model architectures to push the boundaries of what can be achieved in action recognition.

References

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13359–13368.

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Qin Cheng, Jun Cheng, Zhen Liu, Ziliang Ren, and Jianming Liu. 2024. A dense-sparse complementary network for human action recognition based on rgb and skeleton modalities. Expert Systems with Applications, 244:123061.

Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. 2021. UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16266–16275.

Jinfu Liu, Xinshun Wang, Can Wang, Yuan Gao, and Mengyuan Liu. 2024a. Temporal decoupling graph convolutional network for skeleton-based gesture recognition. IEEE Transactions on Multimedia.

Jinfu Liu, Baiqiao Yin, Jiaying Lin, Jiajun Wen, Yue Li, and Mengyuan Liu. 2024b. Hdbn: A novel hybrid dual-branch network for robust skeleton-based action recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshop (ICMEW).

Chiara Plizzari, Marco Cannici, and Matteo Matteucci. 2021. Skeleton-based action recognition via spatial and temporal transformer networks. Computer Vision and Image Understanding, 208:103219.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In CVPR.

Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Q. Miao. 2023. Transformer for skeleton-based action recognition: A review of recent advances. Neurocomputing, 537:164–186.

Sijie Yan, Yuanjun Xiong, Jingbo Wang, and Dahua Lin. 2019. Mmskeleton. https://github.com/open-mmlab/mmskeleton.

Qipeng Zhang, Tian Wang, Mengyi Zhang, Kexin Liu, Peng Shi, and Hichem Snoussi. 2021. Spatial-temporal transformer for skeleton-based action recognition. In 2021 China Automation Congress (CAC), pages 7029–7034.