ISSN: 2993-8643

Research Article

Engineering: Open Access

A Survey of Recent Graph-Based Methods for Skeleton- Based Action Recognition

Abdur Rahman Siam*

Cybersecurity Technology, Canterbury Christ Church University, Changzhou University, China

***Corresponding Author**
Abdur Rahman Siam, Cybersecurity Technology, Canterbury Christ Church University, Changzhou University, China.

Submitted: 2026, Jan 03; **Accepted**: 2026, Jan 30 **Published**: 2026, Feb 06

Citation: Siam, A. R. (2026). A Survey of Recent Graph-Based Methods for Skeleton- Based Action Recognition. *Eng OA, 4*(2), 01-07.

Abstract

Skeleton-based action recognition (SBAR) leverages 3D joint trajectories to recognize human activities while offering privacy, robustness to illumination/background changes, and computational efficiency. Recent progress is dominated by spatio-temporal graph neural networks (ST-GNNs) that model the human body as a graph and learn data-adaptive connectivity, hierarchical structure, compact representations, multimodal supervision, and efficient temporal fusion. This survey focuses on five representative methods CTR-GCN, HD-GCN, InfoGCN, Language Supervised Training (LST), and Temporal Channel Aggregation (TCA-GCN) and positions them within the broader SBAR literature. We analyze modeling assumptions, architectural choices, training objectives, and empirical results on NTU RGB+D 60/120 and Northwestern- UCLA. We additionally contextualize the trajectory from dynamic topology learning to emerging foundation and sequence models reported in 2024-2025. Finally, we summarize open challenges and provide research directions for scalable, robust, and semantically grounded SBAR.

Keywords: Skeleton-Based Action Recognition, Graph Convolutional Networks, Dynamic Topology, Hierarchical Graphs, Information Bottleneck, Language Supervision, Temporal-Channel Aggregation

1. Introduction

Human action recognition aims to infer semantic activities from observations of body motion and has applications in assistive living, human–computer interaction, robotics, surveillance, sports analytics, and clinical assessment. Among sensing modalities, 3D skeleton sequences provide a compact, view-aware representation that abstracts away appearance, background clutter, and lighting. Skeletons are typically captured from depth sensors or multi-view pose estimation pipelines, representing the body as a set of joints connected by bones over time. This structured representation naturally suggests a graph formulation, enabling models to exploit anatomical connectivity while learning action-dependent relations. Spatio-temporal graph neural networks (ST-GNNs) have become the dominant paradigm for SBAR. They treat joints as graph nodes, bones as edges, and apply graph convolutions (spatial) jointly with temporal convolutions/attention (temporal). However, standard fixed graphs cannot capture action-specific correlations (e.g., hands interacting with objects) and may under-utilize multi-scale dependencies. The five focal papers address these limitations via dynamic topology learning (CTR-GCN), hierarchical decomposition (HD-GCN), information-theoretic representation learning (InfoGCN), language-based supervision (LST), and temporal–channel feature aggregation (TCA-GCN) [1,2].

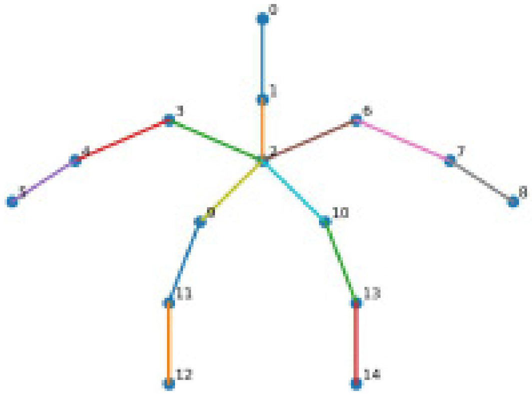

Figure 1: Example Skeleton Graph (nodes=joints, edges=bones). This survey uses Abstracted Illustrations, not Reproduced Paper Figures. (Author-generated Conceptual Diagram.)

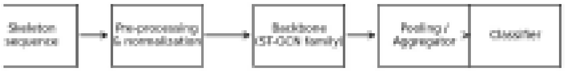

Figure 2: Typical SBAR Pipeline: Skeleton Sequence → Normalization → Spatio-Temporal Backbone → Aggregation → Classification. (Author-Generated Conceptual Diagram.)

2. Background and Problem Formulation

2.1 Skeleton Representation

A skeleton sequence is commonly represented as X ∈ R^{T×V×C}, where T is the number of frames, V the number of joints, and C the coordinate dimension (often 3). Beyond raw joints, many systems construct complementary modalities: bones (vector between connected joints), joint motion (temporal differences), and bone motion. Multi-stream fusion of these modalities is standard in state-of-the-art pipelines.

2.2 Spatio-Temporal Graph Learning

Given an adjacency matrix A encoding joint connectivity, spatial graph convolutions propagate features across joints. Temporal modeling (e.g., temporal convolutions, attention, or recurrent units) captures dynamics across frames. Key research questions include: (1) how to define or learn A (static vs. dynamic, global vs. channel-wise),
(2) how to capture multi-scale and long-range dependencies, (3) how to learn compact but discriminative representations, and (4) how to incorporate auxiliary supervision (text, contrastive objectives) while maintaining efficient inference.

3. Benchmark Datasets and Evaluation Protocols

We summarize common SBAR benchmarks used by the focal works. NTU RGB+D 60 contains 60 classes and 56,880 samples with cross-subject (X-Sub) and cross- view (X-View) protocols. NTU RGB+D 120 extends it to 120 classes and 114,480 samples, typically evaluated using cross-subject (X-Sub) and cross-setup (X-Set). Northwestern-UCLA is a smaller multi-view dataset with 10 classes and 1,494 samples, evaluated by training on two views and testing on the third.

Dataset	#Class es	#Sampl es	#Joint s	Protoco ls	Notes
NTU	60	56,880	25	X-Sub,	3 Kinect v2
RGB+D 60				X-View	cameras;
					multi-modal
					(RGB,
					depth, IR,
					skeleton)
NTU	120	114,480	25	X-Sub,	Adds 60
RGB+D 120				X-Set	classes; 106
					subjects;

					more views/setups
Northwestern-UCLA	10	1,494	20	Cross-view	3 views; small-scale; strong view generalization on test
Kinetics-Skeleton	400	~300k	18	Top-1/Top-5	Large-scale 2D skeletons derived from

Table 1: Common SBAR Datasets and Evaluation Protocols (Counts Reflect Standard Public Dataset Descriptions)

4. Taxonomy of the Focal Methods

The five focal approaches can be organized along complementary axes: (1) topology learning (how the adjacency is adapted to an input sequence and/or feature channels), (2) hierarchical structure (multi-level decomposition into parts and subgraphs), (3) representation objectives (information bottleneck, compactness, disentanglement), (4) supervision signals (labels only vs. language/text guidance), and (5) fusion/aggregation strategies (how temporal and channel information is combined efficiently).

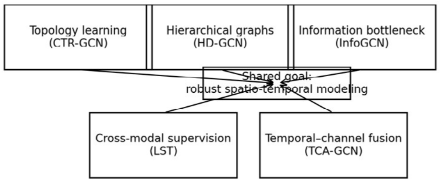
Topology learning (CTR-GCN)
Hierarchical graphs (HD-GCN)
Information bottleneck (InfoGCN)
Shared goal: robust spatio-temporal modeling
Cross-modal supervision (LST)
Temporal–channel fusion (TCA-GCN)

Figure 3: High-Level Taxonomy of the Five Focal Methods and the Design Axis each Emphasizes

5. Detailed Review of Five Representative Approaches

5.1 CTR-GCN (Channel-wise Topology Refinement)

CTR-GCN introduces channel-wise topology refinement, learning feature-dependent adjacency matrices per channel group to capture flexible correlations beyond fixed anatomy. The refinement formulation unifies prior GCN variants and relaxes constraints that limit expressive power in earlier designs. In practice, CTR-GCN typically uses a multi-stream setup (joint, bone, joint motion, bone motion), fusing stream scores at inference.

5.2 HD-GCN (Hierarchically Decomposed Graphs)

HD-GCN proposes an HD-Graph that decomposes the skeleton into a hierarchy of semantically meaningful subgraphs and progressively composes them. This hierarchical decomposition promotes multi-scale reasoning: local part-level dynamics and global body coordination. Compared with flat graph learning, the explicit hierarchy provides inductive bias for structured dependency modeling.

5.3 InfoGCN (Information Bottleneck + Attention-based GCN)

InfoGCN frames SBAR as representation learning under an information bottleneck objective: learn latent codes that are maximally predictive of action labels while being compact. It combines an information-theoretic loss with attention-based graph convolution that captures context- dependent intrinsic topology. The method also proposes an encoding mechanism for stable and discriminative latent representations.

5.4 LST (Language Supervised Training)

LST corresponds to the Generative Action Description Prompts (GAP) framework, which uses a large language model to generate global/local action-part descriptions and aligns skeleton representations with text during training. LST augments skeleton-action supervision with natural language descriptions of actions. Language provides semantic priors (e.g., which body parts move, interaction attributes) that are not explicit in one-hot labels. During training, an auxiliary language-guided objective aligns skeleton features with textual embeddings, improving generalization while keeping inference unchanged.

5.5 TCA-GCN (Temporal–Channel Aggregation)

TCA-GCN targets efficient fusion between temporal modeling and channel-wise spatial topology learning. It proposes a temporal aggregation module for temporal dependencies and a channel aggregation module that combines spatial channel-wise topological features with temporal features. The design aims to avoid overemphasizing only spatial or only temporal cues and to improve multi-scale temporal feature extraction.

6. Empirical Comparison and Discussion

Table II summarizes the key architectural choices of the five methods. Table III reports their headline Top-1 accuracies on the major benchmarks using the configurations reported in the original papers ([3]–[7]) (often multi-stream ensembles). Because ensemble size and input modalities differ across works, direct comparison must be interpreted cautiously; we therefore report modality count explicitly.

Method	Core idea	Topology	Temporal modeling	Aux. objective	Typical modalities
CTR-GCN [3]	Channel-wise topology refinement	Dynamic, channel-wise	TCN + GCN blocks	-	4-stream (J,B,JM, BM)
HD-GCN [4]	Hierarchically decomposed graphs	Hierarchical, learnable	GCN in hierarchy + TCN	-	Multi-stream variants
InfoGCN [5]	Info bottleneck + attention GCN	Dynamic, attention-based	GCN + TCN	Information bottleneck	Often 6-stream ensemble
LST (GAP) [6]	Text-guided training for semantics	Backbone-dependent	Backbone-dependent	Language alignment loss	Standard multi-stream; training-time only (no extra inference cost)
TCA-GCN [7]	Temporal–channel aggregation	Dynamic spatial & temporal	Temporal aggregation module	-	4-stream (J,B,JM, BM)

Table 2: Summary of Modeling Choices in the Five Focal Methods

Method	Streams	NTU60 X-Sub	NTU60 X-View	NTU120 X-Sub	NTU120 X-Set	NW-UCLA
CTR-GCN [3]	4	92.4	96.8	88.9	90.6	96.5
HD-GCN [4]	(reported)	93.4	97.2	90.1	91.6	97.0
InfoGCN [5]	6 (ensemble)	93.0	97.1	89.8	91.2	97.0
LST (GAP) [6]	4 (ensemble)	92.9	97.0	89.9	91.1	97.2
TCA-GCN [7]	4 (ensemble)	92.8	97.0	89.4	90.8	96.8

Table 3: Headline Top-1 Accuracies (%) Reported by the Focal Papers on Standard Benchmarks

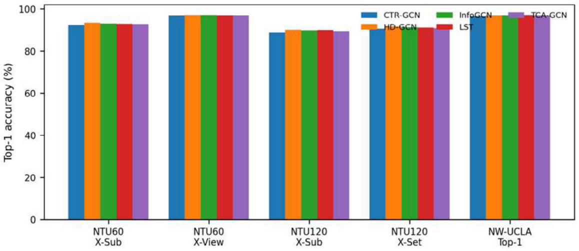

Figure 4: Benchmark Accuracies of the Five Focal Methods (as Reported in their Papers)

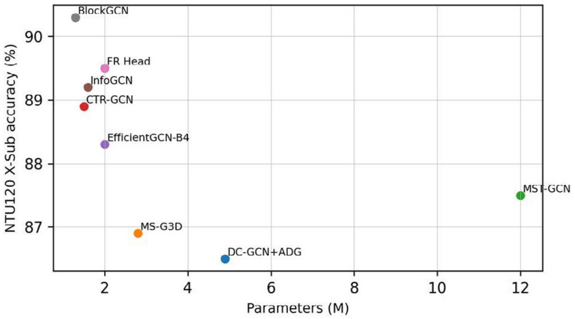

Figure 5: Accuracy–Parameter Trade-Off on NTU120 X-Sub using 4-Modality Results Reported in Block GCN (CVPR 2024)

6.1 What Changes across the Five Methods?

CTR-GCN emphasizes adaptive topology per feature channel, improving expressiveness with minimal parameter overhead. HD-GCN emphasizes structured multi-scale composition, enabling explicit part-to-whole reasoning. InfoGCN introduces an information-theoretic objective that encourages compact representations and uses attention to infer intrinsic topology. LST leverages language to inject semantic priors during training and is particularly relevant for low-shot or compositional generalization settings. TCA-GCN focuses on efficient integration of temporal and channel-wise spatial cues to avoid bias toward either dimension.

6.2 Reproducibility and Fairness Considerations

Reported results in SBAR are sensitive to input modalities (joint/bone/motion), ensemble strategy (number of streams), training schedules, and preprocessing (normalization, velocity computation, sampling rate). Recent works explicitly re-evaluate prior methods under unified 4-modality settings to improve fairness (e.g., BlockGCN, CVPR 2024). We recommend that future papers report single-stream (joint-only) results alongside multi-stream ensembles, disclose FLOPs/params, and release code for standardized comparison [8-10].

7. Recent Directions (2024–2025) and Emerging Trends

Since 2024, SBAR research has increasingly explored (1) stronger topology-aware GCN variants (e.g., BlockGCN, CVPR 2024), (2) hybrid architectures combining GCNs with attention/transformers, and (3) broader supervision signals, including text prompts and foundation-model pretraining for pose-based tasks. At the same time, sequence-model alternatives such as Mamba-style state space models have begun to appear in the SBAR literature, motivated by efficient long-range modeling. Recent work has also expanded toward zero-shot SBAR via explicit language–skeleton alignment (e.g., PURLS, CVPR 2024) and efficient long-sequence modeling with selective state space models (e.g., Simba, arXiv 2024) [11,12].

These trends suggest a convergence toward: (1) scalable pretraining on large motion corpora, (2) modality alignment between skeleton and language for zero-shot recognition, (3) efficiency-aware design for on-device deployment, and (4) robustness to view/domain shifts and noisy pose estimation.

8. Open Challenges and Future Research Directions

• Robustness to noisy or incomplete skeletons: occlusions, missing joints, and pose-estimation artifacts remain major failure modes.

• Generalization: cross-view, cross-subject, and cross- domain transfer (camera, environment, sensor) require stronger invariances and adaptation methods.

• Data efficiency: few-shot and zero-shot SBAR can benefit from language supervision, contrastive pretraining, and synthetic augmentation.

• Structured interpretability: learning which joints/parts and which time segments drive predictions is critical for clinical and safety applications.

• Unified evaluation: standardized protocols for modality/ensemble reporting, compute metrics, and ablations are needed to reduce benchmark inflation.

• Beyond classification: segmentation, early action prediction, and interaction understanding with multiple people and objects are active frontiers.

9. Conclusion

This survey reviewed five representative SBAR methods that illustrate the modern design space of graph-based skeleton modeling: channel-wise dynamic topology refinement (CTR-GCN), hierarchical decomposition (HD- GCN), information-theoretic representation learning (InfoGCN), language-supervised training (LST), and temporal–channel aggregation (TCA-GCN). Across standard benchmarks, all five demonstrate that learning beyond static anatomy—via adaptive topology, hierarchy, compact objectives, or semantic supervision— consistently improves recognition. Looking ahead, the most promising directions are foundation-model pretraining for pose, cross-modal alignment with language, and efficiency-robustness trade-offs necessary for real-world deployment.

References

1. Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
2. Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12026-12035).
3. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., & Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13359-13368).
4. Lee, J., Lee, M., Lee, D., & Lee, S. (2023). Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10444-10453).
5. Chi, H. G., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., & Ramani, K. (2022). Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20186-20196).
6. Xiang, W., Li, C., Zhou, Y., Wang, B., & Zhang, L. (2023). Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10276-10285).
7. Wang, S., Zhang, Y., Zhao, M., Qi, H., Wang, K., Wei, F., & Jiang, Y. (2022). Skeleton-based action recognition via temporal-channel aggregation. *arXiv preprint arXiv:2205.15936.*
8. Zhou, Y., Yan, X., Cheng, Z. Q., Yan, Y., Dai, Q., & Hua, X. S. (2024). Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2049-2058).

9. Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1010-1019).
10. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2019). Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence, 42*(10), 2684-2701.
11. Chaudhuri, S., & Bhattacharya, S. (2024). Simba: Mamba augmented u-shiftgcn for skeletal action recognition in videos. *arXiv preprint arXiv:2404.07645.*
12. Zhu, A., Ke, Q., Gong, M., & Bailey, J. (2024). Part-aware unified representation of language and skeleton for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18761-18770).