SCIENCE CHINA
Information Sciences

• LETTER •

June 2026, Vol. 69, Iss. 6, 169102:1–169102:2
https://doi.org/10.1007/s11432-025-4748-1

AS-CAR: adaptive topology evolution with semantic alignment for continual action recognition

Xingyu ZHU[1], Xiangbo SHU[1*], Binqian XU[1], Liyan ZHANG[2] & Jinhui TANG[3]

[1]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
[2]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*
[3]*College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China*

Received 6 July 2025/Revised 10 November 2025/Accepted 22 December 2025/Published online 29 January 2026

Citation Zhu X Y, Shu X B, Xu B Q, et al. AS-CAR: adaptive topology evolution with semantic alignment for continual action recognition. Sci China Inf Sci, 2026, 69(6): 169102, https://doi.org/10.1007/s11432-025-4748-1

Few-shot class-incremental action recognition [1] (FSCAR) on skeletal data aims to enable models to continually recognize new actions from only a few samples while retaining knowledge of previously learned ones. Unlike static image classification, skeleton-based FSCAR must handle dynamic spatiotemporal structures, where human joints exhibit fine-grained and stage-dependent topology variations over time. Although skeleton data offer compact and background-robust representations, these topological variations present two major challenges. First, neglecting stage-specific skeletal structures limits the model's ability to persistently capture evolving human motion patterns. Second, inconsistent graph topologies across stages blur class boundaries between base and incremental actions, exacerbating few-shot classification errors.

Effective FSCAR therefore requires preserving and adapting stage-specific skeletal topologies to maintain continuity of motion understanding, while simultaneously mitigating cross-stage structural inconsistency. Motivated by these observations, our work introduces an FSCAR framework (AS-CAR) that enables continual knowledge evolution and robust class separation under few-shot incremental settings. AS-CAR incorporates two key components. First, an adaptive topology evolution module learns stage-specific skeletal graph structures, enabling the model to adaptively preserve and extend motion representations across incremental stages. Second, a semantic structure transfer mechanism constructs class-anchor graphs by transferring semantic topology from pretrained language models, refining decision boundaries via optimal transport to alleviate few-shot overfitting. Together, these components enable AS-CAR to achieve robust continual learning and effective cross-stage knowledge transfer under few-shot conditions. The conceptual framework of AS-CAR is illustrated in Figure 1.

Problem definition. FSCAR aims to continually learn new action categories from a few examples while retaining performance on previously learned ones [2]. The task follows a multi-stage protocol. (1) Base session ($\mathcal{D}^0$): the model is trained on abundant data with base classes $\mathcal{C}^0$ to learn general action representations. (2) Incremental sessions ($\mathcal{D}^t, t = 1, \ldots, T$): at each stage t, the model receives N novel classes in a k-shot setting: $\mathcal{D}^t = \{(x_i, y_i, l_i)\}_{i=1}^{kN}$, $x_i \in \mathcal{X}^t$, $y_i \in \mathcal{Y}^t$, $l_i \in \mathcal{C}^t$, where x_i is a skeletal video sequence, y_i is the class label, and l_i is its descriptor (e.g., class name).

Backbone layer. Given a skeletal video sample $x \in \mathbb{R}^{F\times V\times S}$, where F is the number of frames, V the number of joints, and S the number of spatial coordinate channels, we construct a dynamic human skeleton graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ denoting the set of joints (nodes) and $\mathcal{E}$ the set of inter-joint connections (edges). Spatiotemporal features are extracted through an L-layer graph convolutional network [3] (GCN). The propagation of hidden features from layer l to layer l+1, denoted by $H^{(l)} \to H^{(l+1)}$, is defined as

$$H^{(l+1)} = \sum_{s=1}^{S} \sigma\big(A_s^{(l+1)} H^{(l)} M_s^{(l+1)}\big), \tag{1}$$

where $A_s^{(l+1)} \in \mathbb{R}^{V\times V}$ represents the adaptive graph-topology matrix for channel s, $M_s^{(l+1)} \in \mathbb{R}^{S^{(l)}\times S^{(l+1)}}$ is a learnable projection matrix, and $\sigma(\cdot)$ denotes the ReLU activation function. Finally, this GCN is coupled with a temporal convolutional module to capture motion dynamics and joint dependencies across time.

Progressive momentum adaptation (PMA). To ensure stable adaptation under few-shot incremental learning, we propose the PMA strategy for the visual encoder $\mathcal{F}_\theta(\cdot)$, which is composed of L hierarchical layers $\{\mathcal{F}_\theta^1(\cdot), \mathcal{F}_\theta^2(\cdot), \ldots, \mathcal{F}_\theta^L(\cdot)\}$. After completing the base session training, the shallow layers $\{\mathcal{F}_\theta^1(\cdot), \ldots, \mathcal{F}_\theta^m(\cdot)\}$ are frozen to preserve the general motion semantics captured during base learning, as these layers primarily encode task-agnostic representations. During subsequent incremental sessions, only the deeper layers $\{\mathcal{F}_\theta^{m+1}(\cdot), \ldots, \mathcal{F}_\theta^L(\cdot)\}$ are updated to enable efficient adaptation to new action classes while preventing interference with previously learned knowledge. However, due to the limited number of samples per class in each incremental session, direct fine-tuning of the unfrozen layers can easily lead to overfitting. To mitigate this, PMA introduces an exponential moving average (EMA) update mechanism to gradually refine the trainable parameters, formulated as

$$\theta_t = \beta\,\theta_{t-1} + (1-\beta)\,\hat{\theta}_t, \tag{2}$$

where $\hat{\theta}_t$ represents the parameters that have been updated

* Corresponding author (email: shuxb@njust.edu.cn)

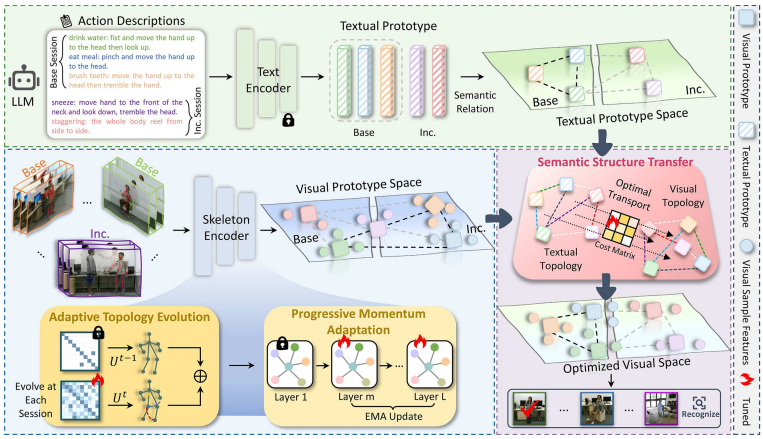

Figure 1 (Color online) Conceptual framework of the proposed AS-CAR method for FSCAR.

through gradient optimization for the current task t, and $\beta \in [0,1]$ is the momentum coefficient controlling temporal smoothness.

Adaptive topology evolution (ATE). In a continual skeleton-based action recognition framework, the evolution of graph topology in each session is essential for maintaining the continuity of human motion representations. To address the heterogeneous nature of skeletal data at the incremental stage t, we introduce the ATE module that enables the model to capture session-specific inductive biases embedded in the skeletal topology, thereby reinforcing persistent understanding of human motion patterns. For each incremental session t, a learnable modulation matrix U^t, having the same shape as the base adjacency matrix A, is learned to represent the session-specific skeletal graph. The modulation is then integrated into the original graph convolution propagation as follows:

$$H^{(l+1)} = \sum_{s=1}^{S} \sigma\left((A_s^{(l+1)} + \mu\, U_s^{t,(l+1)}) H^{(l)} M_s^{(l+1)} \right), \qquad (3)$$

where $U_s^{t,(l+1)}$ denotes the topology modulation matrix for channel s at layer $l{+}1$ in session t, and $\mu = 0.4$ controls the strength of topology adaptation. Throughout the continual learning process, the composite topologies $A + \mu U^t$ from all previous sessions $1, \ldots, t{-}1$ are kept frozen, while only the current session's U^t is updated.

Semantic structure transfer (SST). In FSCAR, unclear boundaries between base and incremental classes often lead to overfitting. To address this, we propose a semantic structure transfer strategy that introduces semantic relations from a pretrained language model into the visual feature space [4]. For each action class, we prompt a large language model to generate detailed motion descriptions $\mathcal{A} = \{\widetilde{\mathcal{C}}_1, \widetilde{\mathcal{C}}_2, \ldots, \widetilde{\mathcal{C}}_C\}$, which are encoded by a frozen text encoder $\mathcal{T}\varphi(\cdot)$ to obtain textual prototypes $P^{\text{text}} = \mathcal{T}\varphi(\mathcal{A}) \in \mathbb{R}^{C \times d_t}$. Corresponding visual prototypes $P^{\text{vis}} \in \mathbb{R}^{C \times d_v}$ are derived from the visual backbone. More details of the prompt process can be found in Appendix C.

We align the pairwise relational graphs of P^{text} and P^{vis} using the entropy-regularized Gromov-Wasserstein [5] (EGW) distance to preserve cross-modal structural consistency. Let D_1 and D_2 be the pairwise distance matrices in text and visual spaces, respectively. The EGW loss is defined as

$$\begin{aligned} \mathcal{L}_{GW} &= \min_{T \in \Pi(p,q)} \sum_{i,j,k,l} \big(D_1[i,k] - D_2[j,l]\big)^2 T_{i,j}\, T_{k,l} \;-\; \epsilon\, H(T) \\ &= \min_{T \in \Pi(p,q)} \sum_{i,j} T_{i,j}\, M_{i,j} \;-\; \epsilon\, H(T), \end{aligned} \qquad (4)$$

where $\Pi(p,q)$ is the set of couplings with marginals $p = q = \frac{1}{C}\mathbf{1}C$, $H(T) = -\sum i,j T_{i,j} \log T_{i,j}$ denotes the entropy term, $M = (D_1 \odot D_1)\, p\mathbf{1}^\top + \mathbf{1}\big((D_2 \odot D_2)\, q\big)^\top - 2 D_1 T D_2^\top$, and ϵ controls regularization strength. This formulation enables differentiable alignment between textual and visual structures, effectively transferring semantic topology across modalities.

Experiments. We conducted experiments on three benchmark datasets, namely, NTU-60, NTU-120, and PKU-MMD I. The proposed method was evaluated against several recently proposed methods. The results showed that AS-CAR achieves state-of-the-art performance in FSCAR. We verified the effectiveness of AS-CAR in handling gradually changing scenarios and its robust anti-forgetting capability regarding source knowledge. Moreover, we conducted ablation studies and hyperparameter analyses to validate the key components of AS-CAR. Visualization studies provide deep insights into AS-CAR. Details of the experimental settings and a comprehensive analysis of the results can be found in Appendixes C–E.

Conclusion. In this study, we tackle the challenges of dynamic topology evolution and semantic drift in few-shot class-incremental skeletal action recognition through the proposed framework. An adaptive topology evolution module dynamically modulates the GCN structure to capture stage-specific motion patterns, while a semantic structure transfer strategy leverages pretrained semantic topology to refine class boundaries. Future work will focus on improving scalability and exploring how to fully leverage the rich knowledge from the base stage to guide learning in the incremental stage.

Acknowledgements The work was supported by National Natural Science Foundation of China (Grant Nos. U25A20442, 62427808).

Supporting information Appendixes A–F. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Garg P, Joseph K, Balasubramanian V N, et al. Poet: prompt offset tuning for continual human action adaptation. In: Proceedings of European Conference on Computer Vision, 2024. 436–455
2 Wang L, Zhang X, Su H, et al. A comprehensive survey of continual learning: theory, method and application. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 5362–5383
3 Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of Conference on Artificial Intelligence, 2018. 7444–7452
4 Park K H, Song K, Park G M. Pre-trained vision and language transformers are few-shot incremental learners. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2024. 23881–23890
5 Han Y, Hui L, Jiang H, et al. Generative subgraph contrast for self-supervised graph representation learning. In: Proceedings of European Conference on Computer Vision, 2022. 91–107