# Skeleton-sequence-based Early Action Recognition by Using Graph Convolutional Neural Networks and Knowledge Distillation Techniques

Wen-Nung Lie[1], Kien Truc Le[1], Veasna Vann[1], Jui-Chiu Chiang[1], Ngoc Dung Bui[2]

[1]Department of Electrical Engineering,
National Chung Cheng University (CCU), Taiwan
E-mail: ieewnl@ccu.edu.tw
[2]Faculty of Information Technology,
University of Transport and Communications (UTC), Hanoi, Vietnam

**Abstract—Early action recognition, a critical task in human behavior analysis, aims to predict the class label of an action before the action is fully executed. While 3D skeleton data provides a compact and robust representation for this task, early prediction remains challenging due to the limited discriminative information available at the initial stages of an action. To address this issue, we propose TS-GCN, a novel Teacher-Student learning architecture based on Graph Convolutional Networks (GCNs). Leveraging the strengths of GCNs in modeling both spatial and temporal dependencies, our approach facilitates effective knowledge transfer from a powerful pre-trained teacher model to a lightweight student model, thereby improving the student's predictive capabilities on partially observed sequences. By integrating high-order joint information, we enhance the distillation process, leading to state-of-the-art performance on benchmark datasets. Experimental results demonstrate that our method achieves superior accuracy in early action recognition, even under limited observational data, also underscoring its potential for real-time applications.**

***Keywords— Teacher-Student learning architecture, action recognition, early action recognition, graph convolutional network.***

## I. INTRODUCTION

In recent years, the rapid advancement of artificial intelligence (AI) and machine learning technologies has significantly influenced various domains, including computer vision tasks, human-computer interaction, and human behavior analysis. A crucial application is skeleton-based action recognition, referring to the prediction of the class label of a human action or gesture (e.g., eating, talking, phoning, etc.) according to the skeleton sequence captured via RGB or depth cameras. A variation of human action recognition (HAR) is called Early Action Recognition (EAR), implying recognizing the label of an action before it is fully executed, or, only partially observed (see Fig. 1). This capability is vital for numerous real-world applications, such as health-care monitoring, security surveillance, autonomous driving, and human-computer interaction. By anticipating actions in advance, potential crises can be mitigated and timely alerts issued, thereby enhancing both safety and operational efficiency.

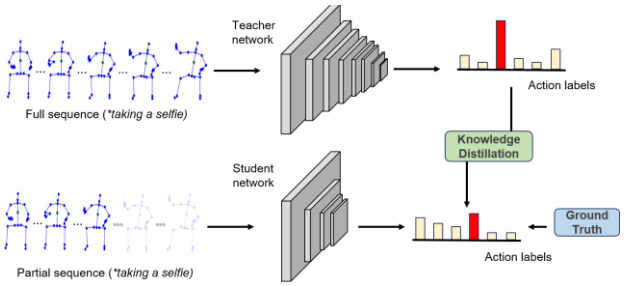


Fig. 1. Teacher-Student framework for early action recognition. This schematic diagram illustrates the rationale behind our proposal to distill knowledge from a traditional action recognition sub-system (top part) to an early action recognition sub-system (bottom part).

Current approaches to HAR primarily rely on two types of input data: 3D skeleton sequences and RGB image sequences. Among these, 3D skeleton data which captures the trajectories of human joints in three-dimensional space certainly offers a more meaningful and robust representation of human movements compared to RGB data. This makes it especially well-suited for EAR tasks, where only partial action sequences are available for analysis.

Despite its advantages, as illustrated in Fig. 1, EAR based on 3D skeleton data remains a challenging task due to the limited and often ambiguous information present in the early stages of an action. To address this challenge, we propose a novel Teacher-Student Learning Architecture based on Graph Convolutional Network (GCN), abbreviated as TS-GCN, for skeleton-based early action recognition. Our method leverages the powerful spatio-temporal modeling capabilities of GCNs and employs a teacher-student learning

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

framework to facilitate knowledge transfer from a well-trained powerful teacher model to a lightweight student model. This transfer enhances the student model's ability to make accurate predictions based on incomplete/partial action sequences.

In our framework, the teacher model serves as a high-capacity guide, providing supervision and refined representations that improve the discriminative power of the student model. Extensive experiments conducted on benchmark datasets validate the effectiveness of our approach, achieving state-of-the-art performance in EAR tasks.

In summary, our contributions are summarized as follows.

1. Novel Architecture Design: We introduce a teacher-student learning framework tailored for GCNs in skeleton-based EAR tasks. This architecture capitalizes on the complementary strengths of both teacher and student models to enhance overall system performance.
2. Innovative Training Paradigm: Our approach features a distinctive training strategy in which the teacher model is pre-trained in an off-line manner. During student network training, the pre-trained teacher model provides guidance, enabling more efficient and effective knowledge transfer. This method not only accelerates the student model's convergence but also improves its generalization.
3. Improved Recognition Accuracy: Through effective GCNs and knowledge distillation, our approach significantly boosts the EAR accuracy. Experimental results demonstrate that our method outperforms existing techniques, setting a new benchmark on public datasets.

## II. Related Work

### A. Traditional Human Action Recognition (HAR)

Skeleton-based HAR has long been a critical research area within human behavior analysis. Due to the easy modeling of a human skeleton as a graph, state-of-the-art (SOTA) methods [1] nearly all rely on the Graph Convolutional Network (GCN) for spatio-temporal feature extraction. The pioneer work for GCN-based approaches was originated in [2], called Spatial-Temporal Graph Convolutional Networks (ST-GCNs), which models spatial and temporal features via multi-layer GCN and TCN (Temporal Convolution Network), respectively. Several extensions based on ST-GCN [2] conducted a multi-stream or multi-ensemble architecture to promote the performance substantially, such as AGCN [3] (2-stream), MS-AAGCN (4-stream) [4], MS-G3D (2-stream) [5], etc. In addition to different joint features extracted for the input of the multi-stream/-ensemble architecture, a learnable adjacency matrix (instead of a fixed graph topology used in ST-GCN [2]) for graph convolution process was also adopted to make the GCN adaptive to different action types. The above models achieve state-of-the-art performances on major skeleton-based action recognition datasets, including NTU RGB+D [6] and Kinetics Skeleton [7].

In [8], Lie *et al.* introduced a novel method that incorporates high-order joint information**,** such as velocity and acceleration, into the GCN input. This enhancement allows the model to better capture complex motion patterns by including up to third-order features. These features are fused into the model in two ways: early fusion and late fusion. Experimental results demonstrate that their approach improved recognition accuracy by 2.55% and 1.32% with respect to RAGCN [9] and MS-AAGCN [4], respectively, on the NTU RGB+D 60 dataset [6].

### B. Early Action Recognition

Early action recognition - predicting the class label of an action or gesture before it is fully executed - has garnered significant interest due to its critical role in real-world applications such as surveillance, autonomous systems, and healthcare. The primary challenge in this task lies in the insufficient discriminative information available from incomplete temporal sequences, making it difficult to distinguish between similar actions in the early stages. To address this issue, various methods have been proposed. For instances, [10] introduced a hardness-guided discrimination network that focuses on hard-to-classify samples to improve early activity prediction. While effective, this focus may lead to overfitting, and the model's complexity can hinder its applicability in real-time settings. Furthermore, identifying the optimal hardness threshold requires extensive hyperparameter tuning, which may vary across datasets.

In another approach, [11] proposed scene-aware spatio-temporal Graph Neural Networks (GNNs) for few-shot early action prediction by leveraging contextual scene information. However, this reliance on static or consistent scene context can limit the model's generalizability in dynamic or heterogeneous environments. Additionally, few-shot learning methods may struggle to differentiate between subtly varying actions without sophisticated feature extraction.

Similarly, [12] employed action-semantic knowledge to align predicted actions with their semantic context, improving coherence in early prediction. However, this method depends on a rich and well-structured semantic knowledge base, which can be difficult to build and maintain. Stergiou and Damen [13] introduced a temporally progressive attention mechanism that aggregates predictions from multiple models to refine early action recognition. While the ensemble strategy improves accuracy, it

significantly increases inference time and computational cost. Moreover, balancing the contributions from various models requires careful attention and fine-tuning.

Dear-Net [14] focuses on capturing diversities in skeleton data to improve early action prediction by modeling variations in action execution. Although effective in theory, the model's complexity leads to high computational overhead, and excessive reliance on diversity may adversely affect prediction accuracy and speed. A different strategy was presented in [15], which uses a policy-based reinforcement learning framework for early action recognition. The method improves performance by selectively excluding certain categories. However, this can limit the model's generalizability and scalability. Additionally, the design of an effective reward function is non-trivial and requires careful crafting to suit complex recognition tasks.

Generative approaches have also been explored. For example, Zhang *et al.* [16] used Generative Adversarial Networks (GANs) to predict unobserved frames, aiming to improve generalizability. However, predicting numerous future frames from limited early observations can introduce significant bias and yield inaccurate predictions. Contrastive learning has recently shown promise in early action recognition. TODO-Net [17] leverages a temporally observed domain contrastive network**,** contrasting observed and unobserved frames to improve prediction. Magi-Net [18] further builds on this idea by introducing a meta-negative contrastive learning network**,** which focuses on negative samples to enhance discrimination between similar early-stage actions. While effective, these models often require extensive negative sample mining and introduce substantial computational costs. InfoGCN++ [19] extended the original InfoGCN [20] model by learning from both current and anticipated future motion to generate a holistic action representation. However, its use of the neural Ordinary Differential Equations (ODEs) to model the continuous evolution of hidden states increases computational complexity.

Recent studies [21][22][23] have explored progressive Teacher-Student learning frameworks**,** where knowledge is gradually distilled from a teacher to a student model over time. Although promising, these methods can be time-intensive, and their effectiveness is highly dependent on the quality of the teacher model.

To overcome this limitation, we propose a novel Teacher-Student framework that leverages high-order joint kinematic features, such as joint velocity and acceleration, extracted as in our prior work [8] to enhance knowledge distillation. This enriched information improves the quality of supervision provided by the teacher model, thereby enhancing the student model's ability to recognize actions at earlier stages with higher accuracy and robustness.

## III. PROPOSED METHOD

### *A. Problem statement*

The primary objective of EAR is to forecast the class label of an action before the action is fully executed. This requires the recognition model to make accurate predictions based on incomplete and often ambiguous information**,** particularly during the early stages of an action. Consequently, prediction accuracy tends to suffer due to the lack of discriminative features. To mitigate this limitation, we propose enhancing the student model's performance by transferring the high-order joint kenematic information from a robust teacher model that has access to the entire action sequence. The teacher model serves as an auxiliary source of supervision, guiding the student model through a knowledge distillation process.

To this end, we design a teacher-student learning framework in which the teacher model, trained with full-sequence data, imparts valuable insights to the student model, which operates on partial observations. In the following sections, we first provide a detailed description of both the teacher and student architectures. We then explain how the distillation mechanism is employed to effectively transfer knowledge from the teacher to the student, thereby improving early prediction accuracy.

### *B. Teacher Network Model*

Our design demonstrates that employing a more powerful teacher model can significantly enhance the performance of the student model which often has a light-weight size to release the hardware/computation loadings in on-line applications. Specifically, we adopt both the AAGCN [4] and MS-G3D [5] models as the possible backbones, advanced GCN architectures that have shown superior capabilities in modeling dynamic skeleton sequences. To further strengthen the teacher model, we integrate the high-order information proposed in our prior work [8] with the AAGCN and MS-G3D frameworks. This combination enables the teacher to extract and convey richer, more discriminative representations, often referred to as "dark knowledge", to the student model through knowledge distillation.

Following the methodology outlined in [8], we first employ a View Adaptation (VA) subnetwork [24] to learn optimal translation and rotation parameters to automatically transform the input skeleton into a viewpoint that is best suited for action recognition. For each transformed skeleton in the sequence, we then compute 5 RICH [8] joint features (including *J*, *E*, *S* in spatial doamin and *D*, *A* in temporal

domain, as illustrated in Fig. 2), which encapsulate position, velocity, and acceleration information, as the inputs to a multi-stream or multi-ensemble (here, 5-straem/ensemble) GCN, whose outputs are then fused using weighted summation to generate the final action classification.

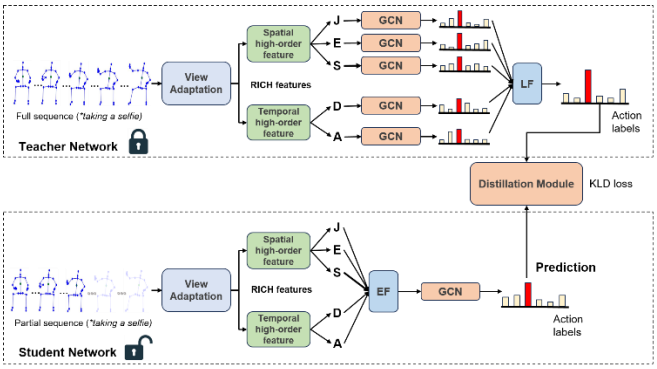


Fig. 2 Our proposed TS-GCN for EAR. The techniques of RICH joint features, VA (view adaptation) module, early fusion, late fusion, and KLD loss were applied.

### *C. Student Network Model*

Though the student model is designed to share a similar architectural backbone as the teacher model, its key distinction lies in the nature of its input - a partially observed action sequence, which reflects the early stages of an ongoing action. To ensure that the student model remains both effective and computationally efficient, we adopt a single stream architecture and an early fusion strategy in [8] for feature integration. The use of early fusion to stack distinct orders of features as different channels in the input allows us to simplify the network model to be a single stream without compromising its predictive capabilities substantially. By aggregating the RICH joint features at an early stage, the model can leverage multi-dimensional skeletal information (e.g., position, velocity, and acceleration) in a compact and unified representation. This not only reduces the computational burden but also preserves critical temporal-spatial correlations necessary for early action recognition.

Overall, this approach strikes a practical balance between accuracy and efficiency**,** making the student model well-suited for real-time and resource-constrained applications, such as surveillance systems**,** robotic interaction, and wearable devices**.**

### *D. TS-GCN – Knowledge Distillation Framework*

As previously discussed, the primary objective of the teacher-student framework is to enable effective knowledge distillation, transferring the representational power of a well-trained teacher model to a student model operating on partial input sequences. To facilitate this, we introduce a Distillation Module (DM), which serves as a bridge between the teacher's output and the student model, as depicted in Fig. 2**.** This module allows the student model to learn not only from ground-truth labels but also from the soft target distributions generated by the teacher, thereby improving its generalization capability and early prediction accuracy.

A widely used loss definition in training a Teacher-Student Network is the Kullback-Leibler Divergence (KLD) loss [25], which quantifies the difference between the softened output distributions from the teacher and student models. These two softened outputs can be calculated using temperature scaling, which helps reveal the relative class similarities encoded in the logits.

Given an output logit vector **z**, the softened probability vector $\boldsymbol{p}^{\tau}$ with a temperature-scaling hyper-parameter τ is defined as:

$$p_i^{\tau} = \frac{exp\left(\frac{z_i}{\tau}\right)}{\sum_{j=1}^{K} exp\left(\frac{z_j}{\tau}\right)}, \tag{1}$$

where $z_i$ is the $i$-th value of $\boldsymbol{z}$, $K$ is the number of classes, $\exp(.)$ is the natural exponential function, and τ >1 controls the smoothness of the distribution.

The distillation loss $L_{disti}$ is then defined as the KLD between the softened teacher output $\boldsymbol{p}^{T,\tau}$ and the student output $\boldsymbol{p}^{S,\tau}$:

$$L_{disti}(\boldsymbol{p}^{S,\tau}, \boldsymbol{p}^{T,\tau}) = \sum_j p_j^{T,\tau} \log \frac{p_j^{T,\tau}}{p_j^{S,\tau}} \tag{2}$$

The use of the temperature parameter τ ensures that the student model benefits from a more informative and smoothed probability distribution, improving its prediction accuracy. This formulation encourages the student model to mimic the class distribution predicted by the teacher, effectively learning from its richer and more informative output, even when the input is incomplete.

We approach the task of early action recognition in a manner consistent with the traditional action recognition methodologies. Accordingly, we define the prediction loss using the standard Cross-Entropy (CE) loss, which encourages the student model to assign high probabilities to the correct class labels based on the available (partial) input sequence. In summary, the total loss function for training our EAR model integrates two key components: the KLD loss and the CE loss:

$$L = L_{CE} + \alpha L_{distill} \tag{3}$$

where α is a hyperparameter that controls the contribution of the distillation loss. Minimizing this total loss allows the student model to effectively inherit the teacher's expertise while refining its own predictive accuracy.

In our teacher-student network, the teacher model is assumed to be pre-trained in advance and off-line manner, meaning it is trained based on the full skeleton sequences and

then fixed/locked for the subsequent learning of the student model. Consequently, our training strategy can be divided into two stages: the first stage involves training of the teacher model on the full sequence dataset and then freezing its parameters when loaded to participate in the training of the student model through the Distillation Module. The second stage is the training phase of the student model, where, with incomplete data and leveraging of the knowledge from the teacher model, the student model is optimized to accomplish the task of early action recognition.

## IV. EXPERIMENTAL RESULTS

We evaluate our method on a 3D skeleton-based action recognition dataset named the NTU RGB+D 60 dataset [6]., We use a GeForce RTX 3090Ti GPU card for training our TS-GCN model, with the Adam optimizer adopted.

The NTU RGB+D 60 dataset [6] is a large-scale RGB, Infrared, and 3D-skeleton-based dataset designed for human action recognition. It contains 56,880 video samples (captured from 3 different viewpoints) of 60 action classes performed by 40 subjects. For the 3D skeleton data, they were extracted from the RGB+D image sequences and form the primary focus for action recognition tasks. Each skeleton contains 25 joints in 3D space and each sequence for training/testing is repeated and extended to have a fixed length of 300 frames (skeletons), as illustrated in Fig. 3(a), which has formed a common protocol for researches based on the NTU RGB+D [6] dataset. There are two standard evaluation protocols outlined by [6]: cross-subject (CS) evaluation and cross-view (CV) evaluation. In the CS setting, 20 subjects are designated for training, while the remaining 20 subjects are reserved for testing. In the CV setting, training is conducted using two viewpoints, and the third viewpoint is used for testing.

In training the student model, videos of different Observation Ratio (OR), which is defined to the proportion in length of the original video clip, will be created. For example, OR=20% means that only the first 20% of frames is remained and the other unobserved 80% will be replaced with zero. The observed and unobserved parts will then be repeated together to fit a total length of 300 frames, as illustrated in Fig. 3(b). Videos of different ORs (here, OR=20%, 40%, 60%, 80%, and 100%) were used as inputs to train a single student model.

We followed the CS testing protocol to evaluate the models on the NTU RGB+D 60 dataset, as illustrated in Table I. It can be observed from Table I that our model achieves an accuracy of 49.14% at OR=0.2, surpassing the previous top-performing Magi-Net model in [18], which stands at 46.68%. At an OR of 0.4, our method's accuracy further increases to 78.12%, once again outperforming the Magi-Net [18], which achieves 75.11%. These results underscore the robustness and effectiveness of our approach in accurately predicting actions with partial sequences, even under limited observed data, establishing a new benchmark in early action recognition.

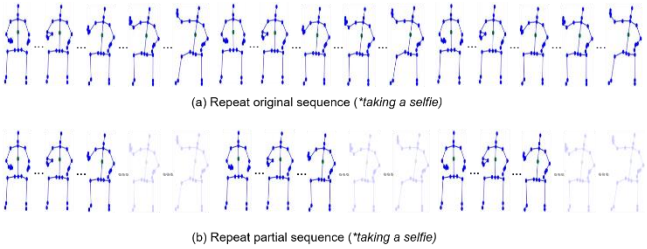


Fig. 3 Repeating frames in preparing the training/testing samples for (a) full or (b) partially observed sequences.

Table II shows an ablation study where the GCN backbones is changed between AAGCN [4] and MS-G3D [5] and when the Teacher-student framework is adopted or not. In both implementations, the original networks are enhanced with 5-stream RICH features (late fusion) in teacher network and 1-stream RICH features (early fusion) in student network. The same teacher-student training method is applied. It can be found that all techniques (early fusion, late fusion, RICH features, KLD loss function, etc.) are able to boost the performances of the baseline networks (i.e., the original AAGCN or MS-G3D) by a large margin, especially at low ORs. This phenomenon can be attributed to the TS-GCN model's augmentation of the student model through the teacher model, such that the “dark knowledge” from the teacher model is more impactful when the student model receives much less information from the input. As more frames are observed, the input actions become more discernible, rendering the “dark knowledge” less effective.

## V. CONCLUSION

This paper has introduced a novel Teacher-Student learning architecture based on Graph Convolutional Networks (GCNs) for early action recognition. By leveraging the strengths of both GCNs and the teacher-student paradigm, our approach effectively transfers knowledge from a pre-trained powerful teacher model to a light-weight student model, enhancing the latter's ability to predict actions from incomplete sequences. Furthermore, the incorporation of high-order joint information further refines the knowledge distillation process, resulting in state-of-the-art performance on benchmark datasets. Our proposed method demonstrates superior accuracy in early action recognition, even with limited observational data, highlighting its potential for real-time applications. Future work could explore the integration of other advanced techniques, such as self-supervised

learning or generative models, to further improve the performance of early action.

Table I. Comparison of recognition accuracy in CS protocol with SOTA methods on the NTU RGB+D 60 dataset. The results are shown with different ORs, ranging from 0.2 to 1.0).

| Methods | Observation Ratios (ORs) | | | | |
|---|---|---|---|---|---|
| | 20% | 40% | 60% | 80% | 100% |
| LGN [26] | 30.04 | 61.78 | 76.14 | 81.57 | 82.64 |
| Local+LGN [26] | 32.12 | 63.82 | 77.02 | 82.45 | 83.19 |
| RL+LSTM [15] | 35.56 | 54.63 | 67.08 | 72.91 | 75.53 |
| BiRNN+GRU [23] | 24.6 | 57.7 | 76.9 | 85.7 | 88.1 |
| BiLSTM+LSTM [21] | 35.85 | 58.45 | 73.86 | 80.06 | 82.01 |
| TemPr [13] | 38.7 | - | - | - | - |
| Local+AGCN-AL [27] | 38.18 | 71.19 | 82.25 | 86.33 | 87.20 |
| GA-Net [28] | 42.53 | 72.64 | 83.12 | 86.75 | 87.21 |
| InfoGCN++ [19] | 44.57 | 73.59 | 81.68 | 84.42 | 85.38 |
| Magi-Net [18] | 46.68 | 75.11 | 84.87 | 88.12 | 88.72 |
| HARDer-Net [10] | 43.22 | 72.43 | 83.17 | 87.00 | 87.80 |
| TODO-Net [17] | 45.95 | 74.37 | 84.61 | 87.71 | 88.62 |
| TS-GCN (Ours) | **49.14** | **78.12** | **86.62** | **89.29** | **89.77** |

Table II. Performance comparison in recognition accuracy (%) with and without our proposed TS-GCN learning when the GCN backbone is varied with 1s-AAGCN [4] and 1s-MS-G3D [5].

| Student Backbone | Methods | Observation Ratios (ORs) | | | | |
|---|---|---|---|---|---|---|
| | | 20% | 40% | 60% | 80% | 100% |
| VA+RICH AAGCN | w/o TS-GCN | 41.58 | 73.41 | 83.93 | 87.38 | 88.38 |
| | w/ TS-GCN | **43.73** | **75.05** | **85.05** | **88.42** | **89.18** |
| | $gain\ \Delta$ | **+2.15** | **+1.64** | **+1.12** | **+0.85** | **+0.80** |
| VA+RICH MS-G3D | w/o TS-GCN | 47.17 | 76.64 | 85.18 | 88.22 | 88.41 |
| | w/ TS-GCN | **49.14** | **78.12** | **86.62** | **89.29** | **89.77** |
| | $gain\ \Delta$ | **+1.97** | **+1.48** | **+1.44** | **+1.07** | **+1.36** |

## REFERENCES


[1] Skeleton Based Action Recognition on NTU RGB+D, https://paperswithcode.com/sota/skeleton-based-action-recognition-on-ntu-rgbd

[2] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *Proc. of the AAAI conf. on Artificial Intelligence* , Vol. 32, No. 1, April 2018.

[3] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 12026-12035, 2019.

[4] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. on Image Processing*, Vol. *29*, 9532-9545, 2020.

[5] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 143-152, 2020.

[6] A. Shahroudy, J. Liu, T.T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010-1019, 2016.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, ... and A. Zisserman, "The kinetics human action video dataset, *arXiv preprint arXiv:1705.06950*, 2017.

[8] W.N. Lie, Y.J. Huang, J.C. Chiang, and Z.Y. Fang, "High-Order Joint Information Input For Graph Convolutional Network Based Action Recognition," *2021 IEEE Int'l Conf. on Image Processing (ICIP)*, pp. 1064-1068, 2021

[9] Y. Song, Z. Zhang, C. Shan, and L. Wang, "Richly Activated Graph Convolutional Network for Robust Skeleton-based Action Recognition," *IEEE Trans. On Circuits and Systems for Video Technology (TCSVT)*, pp.1915-1925, Nov. 2020.

[10] T. Li, Y. Luo, W. Zhang, L. Duan, and J. Liu, "HARDer-Net: Hardness-Guided Discrimination Network for 3D Early Activity Prediction," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 34, No. 12, pp. 12112-12126, 2024.

[11] Y. Hu, J. Gao, and C. Xu, "Learning scene-aware spatio-temporal GNNs for few-shot early action prediction," *IEEE Trans. on Multimedia*, vol. *25*, pp. 2061-2073, 2022.

[12] X. Liu, J. Yin, D. Guo, and H. Liu, "Rich Action-Semantic Consistent Knowledge for Early Action Prediction," *IEEE Trans. on Image Processing*, vol. 33, pp. 479-492, 2024.

[13] A. Stergiou, and D. Damen, "The wisdom of crowds: Temporal progressive attention for early action prediction," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14709-14719, 2023.

[14] R. Wang, J. Liu, Q. Ke, D. Peng, and Y. Lei, "Dear-net: Learning diversities for skeleton-based early action recognition," *IEEE Trans. on Multimedia*, vol. *25*, 1175-1189, 2021.

[15] J. Weng, X. Jiang, W.L. Zheng, and J. Yuan, "Early action recognition with category exclusion using policy-based reinforcement learning," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. *30, no.* 12, pp.4626-4638, 2020.

[16] H.B. Zhang, W. X. Pan, J.X. Du, Q. Lei, Y. Chen, and J.H. Liu, "Adversarial Attention Networks for Early Action Recognition, "*IEEE Trans. on Emerging Topics in Computational Intelligence*, vol. 9, no. 2, pp.1581-1594, 2024.

[17] W. Wang, F. Chang, C. Liu, B. Wang, and Z. Liu, "TODO-Net: Temporally Observed Domain Contrastive Network for 3-D Early Action Prediction," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6122-6133, April 2025.

[18] W. Wang, F. Chang, J. Zhang, R. Yan, C. Liu, B. Wang, and M.Z. Shou, "Magi-net: Meta negative network for early activity prediction," *IEEE Trans. on Image Processing*, vol. *32*, pp. 3254-3265, 2023.

[19] S. Chi, H.G. Chi, Q. Huang, and K. Ramani, "InfoGCN++: Learning Representation by Predicting the Future for Online Skeleton-based Action Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 514-528, 2025.

[20] H.G. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 20186-20196, 2022.

[21] X. Wang, J.F. Hu, J.H. Lai, J. Zhang, and W.S. Zheng, "Progressive teacher-student learning for early action prediction," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3556-3565, 2019.

[22] Y. Cai, H. Li, J.F. Hu, and W.S. Zheng, "Action knowledge transfer for action prediction with partial videos," *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 8118-8125, Jan. 2019.

[23] V. Tran, N. Balasubramanian, and M. Hoai, "Progressive knowledge distillation for early action recognition," *2021 IEEE Int'l Conf. on Image Processing (ICIP)*, pp. 2583-2587, Sept. 2021.

[24] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. *41, no.* 8, pp. 1963-1978, 2019.

[25] T. Kim, J. Oh, N. Kim, S. Cho, and S.Y. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation, *arXiv preprint arXiv:2105.08919,* 2021.

[26] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Learning latent global network for skeleton-based action prediction," *IEEE Trans. on Image Processing*, vol. *29*, pp. 959-970, 2019.

[27] G. Li, N. Li, F. Chang, and C. Liu, "Adaptive graph convolutional network with adversarial learning for skeleton-based action prediction," *IEEE Trans. on Cognitive and Developmental Systems*, vol. *14*, no. 3, pp. 1258-1269, 2021.

[28] W. Wang, F. Chang, C. Liu, G. Li, and B. Wang, "Ga-net: a guidance aware network for skeleton-based early activity recognition," *IEEE Trans. on Multimedia*, vol. *25*, pp. 1061-1073, 2021.