Multi-Modal Graph Convolutional Network with Sinusoidal Encoding for Robust Human Action Segmentation

Hao Xing[1], Kai Zhe Boey[1], Yuankai Wu[2], Darius Burschka[3], Gordon Cheng[1]

***Abstract*— Accurate temporal segmentation of human actions is critical for intelligent robots in collaborative settings, where a precise understanding of sub-activity labels and their temporal structure is essential. However, the inherent noise in both human pose estimation and object detection often leads to over-segmentation errors, disrupting the coherence of action sequences. To address this, we propose a Multi-Modal Graph Convolutional Network (MMGCN) that integrates low-frame-rate (e.g., 1 fps) visual data with high-frame-rate (e.g., 30 fps) motion data (skeleton and object detections) to mitigate fragmentation. Our framework introduces three key contributions. First, a sinusoidal encoding strategy that maps 3D skeleton coordinates into a continuous sin-cos space to enhance spatial representation robustness. Second, a temporal graph fusion module that aligns multi-modal inputs with differing resolutions via hierarchical feature aggregation, Third, inspired by the smooth transitions inherent to human actions, we design SmoothLabelMix, a data augmentation technique that mixes input sequences and labels to generate synthetic training examples with gradual action transitions, enhancing temporal consistency in predictions and reducing over-segmentation artifacts.**

Extensive experiments on the Bimanual Actions Dataset, a public benchmark for human-object interaction understanding, demonstrate that our approach outperforms state-of-the-art methods, especially in action segmentation accuracy, achieving F1@10: 94.5% **and F1@25:** 92.8%**.**

I. INTRODUCTION

Human action segmentation, the task of temporally decomposing continuous activities into coherent sub-action units, is a cornerstone of intelligent robotic systems operating in collaborative environments. Beyond recognizing discrete actions, the systems must understand their temporal structure.

Our preliminary studies [1] on human-robot collaboration reveal that encoder-decoder graph convolutional networks excel at parsing spatiotemporal information by representing human skeleton and object center points as graph nodes. Through attention mechanisms, these networks dynamically update the relationships between nodes, capturing the evolving interactions between human and objects. However, joint misdetections and tracking errors in skeleton data often lead to over-segmentation, fragmenting continuous actions into disjointed segments and disrupting the temporal coherence of action sequences. This limitation becomes particularly critical in human-object interaction scenarios, where both human pose estimation (skeleton) and object detections (bounding boxes) introduce noise. Object misdetections, such as missing and inaccurately localized bounding boxes, lead to inconsistencies in action segmentation.

Recent advances in multi-modal networks [2], [3] have demonstrated promise in addressing noise-induced over-segmentation by integrating structured 3D motion data (e.g., skeletal trajectories and object centroids) with pixel-based visual streams. However, integrating pixel-based RGB features with 3D position information presents significant challenges. The two modalities exhibit divergent characteristics: 3D position data are spatially sparse and structured, while image features are spatially dense and unstructured. Directly concatenating or aligning these representations leads to feature incompatibility. To bridge this gap, we introduce a sinusoidal joint encoding strategy that projects 3D coordinates into a continuous sin-cos space. This transformation downscales noisy joint trajectories while preserving spatial relationships, enabling seamless fusion with visual features.

A second challenge arises from computational efficiency: processing full-frame video data at high temporal resolutions is prohibitively expensive for real-time systems, while naively fusing features across different temporal resolutions degrades performance. A common approach is to upsample low-frequency image features to match the motion sequence's time scale and merge them using a late fusion strategy [3]. However, this final fusion step often results in feature loss due to the mismatch in the fineness of features, and it overlooks fine-grained temporal dependencies and misaligns crucial visual and motion cues. To address this, we propose a Multi-Modal Graph Convolutional Network (MMGCN) that processes high-frequency motion data via a hierarchical graph convolutional network and strategically integrates high-frequency motion features into low-frequency (1 fps) visual cues at a middle stage, ensuring a more coherent fusion with processed motion information later in the pipeline. By processing the two modalities in different temporal resolutions, our method effectively preserves fine-grained temporal dependencies and aligns visual motion cues efficiently.

Finally, inspired by the observation that human actions transition smoothly over time, we develop SmoothLabelMix, a novel data augmentation technique that enforces temporal consistency by smoothing action labels along the temporal dimension and mixing input action sequences and labels in the spatial space.

Overall, the technical contributions of the paper are:

- Sinusoidal encoding: A spatial encoding method that maps 3D skeleton coordinates into a continuous sin-cos space, enabling effective fusion with visual features.

Authors are with [1]Institute for Cognitive Systems, [2]Chair of Media Technology, [3]Machine Vision and Perception Group, School of Computation, Information and Technology, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany. hao.xing@tum.de, kaizhe.boey@tum.de, yuankai.wu@tum.de, burschka@cs.tum.edu, gordon@tum.de

- Multi-Modal Graph Convolutional Network: A hybrid graph-based architecture that processes 30 fps motion sequence through a graph convolutional backbone and integrates 1 fps visual data into skeleton sequences through a parallel branch, balancing computational efficiency and segmentation accuracy.
- SmoothLabelMix Augmentation: A data augmentation strategy combining label smoothing and weighted mixing to align inputs and labels bidirectionally. This reduces over-segmentation artifacts, improves generalization to ambiguous motion boundaries, and enhances robustness to noise.

The remainder of this paper is organized as follows: Section II reviews related work, Section III details the Multi-Modal Graph Convolutional Network architecture, Section IV introduces the SmoothLabelMix data augmentation technique, Section V presents experimental results, and Section VI concludes with future directions.

II. RELATED WORK

In this section, we briefly review three key areas related to our work: Human Action Segmentation, Multi-Modal Models, and Data Augmentation.

A. Human Action Segmentation

Human Action Segmentation, which aims to temporally localize and classify sub-actions within continuous sequences, has evolved from early template-based methods [4], [5] to modern image-based learning architectures [6], [7]. Temporal Convolutional Networks (TCNs) [6] pioneered frame-wise prediction using dilated convolutions, while MS-TCN [7] extended this with multi-stage refinement. Recent works like Action Segment Refinement Framework [8] combine boundary detection and classification to improve temporal localization. Despite progress, these methods suffer from high computational costs, sensitivity to background clutter, and the inclusion of redundant visual information, making them inefficient for real-time applications.

To address these challenges, graph-based networks are considered as a more efficient alternative by modeling human motion as structured spatial and temporal graphs, significantly reducing background noise and computational overhead. Early works, such as Spatial-Temporal Graph Convolutional Network (ST-GCN) [9], introduced graph convolutional networks to sequentially capture joint-level motion patterns across both spatial and temporal dimensions. Subsequent improvements, like Two-Stream Adaptive Graph Convolutional Network (2s-AGCN) [10] and Multi-Scale Graph Network (MS-G3D) [11], incorporated attention mechanisms and multi-scale feature extraction to enhance action recognition. More recently, methods such as Pyramid Graph Convolutional Network (PGCN) [12] and Temporal Fusion Graph Convolutional Network (TFGCN) [1] have explored temporal modeling techniques to further improve segmentation accuracy.

However, despite these advancements, graph-based methods remain highly susceptible to noise from skeleton misdetections, leading to over-segmentation and fragmented action sequences. When combined with noisy object detections in human-object interaction scenarios, these errors further degrade segmentation quality, highlighting the need for robust multi-modal solutions.

B. Multi-Modal Models

Multi-modal learning is a promising method for action recognition, leveraging complementary strengths of visual (RGB/depth) and skeletal modalities to overcome individual limitations. Early works, like [13], [14] fuse images/depths and skeletons per frame to improve the recognition results. Recent advancements focus on intermediate fusion strategies to better align cross-modal features. Kim *et al.* [15] proposed a cross-modal transformer that attends to spatial correspondences between skeleton joints and image regions, improving recognition consistency by dynamically weighting discriminative features. Similarly, Wang *et al.* [16] integrates skeleton joints with RGB patches using a late fusion technique, but its reliance on dense pixel processing limits scalability. To address efficiency, PoseC3D [2] encodes skeletons as pseudo-3D heatmaps and fuses them with RGB features via 3D convolutions, achieving a balance between accuracy and speed. More recently, Tani Hiroaki [3] introduced a temporal attention mechanism that selects one representative image for the entire motion sequence capturing the minimal appearance features necessary for action recognition. While this approach is effective for short-term action recognition, it becomes computationally intensive for long-duration actions.

Building on advancements in multi-modal action recognition, our work extends these principles to human action segmentation, a task requiring precise temporal localization of sub-activities. While existing multi-modal models excel at classification, they often neglect the computational and temporal alignment challenges inherent to segmentation.

C. Data Augmentation

Data augmentation plays a pivotal role in improving the generalization and robustness of temporal models, particularly for action recognition and segmentation tasks. Traditional techniques, such as spatial augmentations (e.g., cropping, flipping) [17] and temporal perturbations (e.g., frame skipping, time warping) [18], focus on enhancing spatial invariance or simulating temporal variations. While effective for classification, these methods often disrupt the temporal coherence required for precise segmentation, where smooth transitions between sub-actions are critical.

Recent approaches adapt image-domain augmentation strategies to temporal tasks. Mixup [19] and CutMix [20], which linearly interpolate inputs and labels, have been extended to videos by blending random clips (e.g., VideoMix [21]). However, these methods prioritize spatial or short-term temporal invariance, inadvertently introducing abrupt label transitions that exacerbate over-segmentation artifacts. For instance, TemporalMix [22] downsamples two

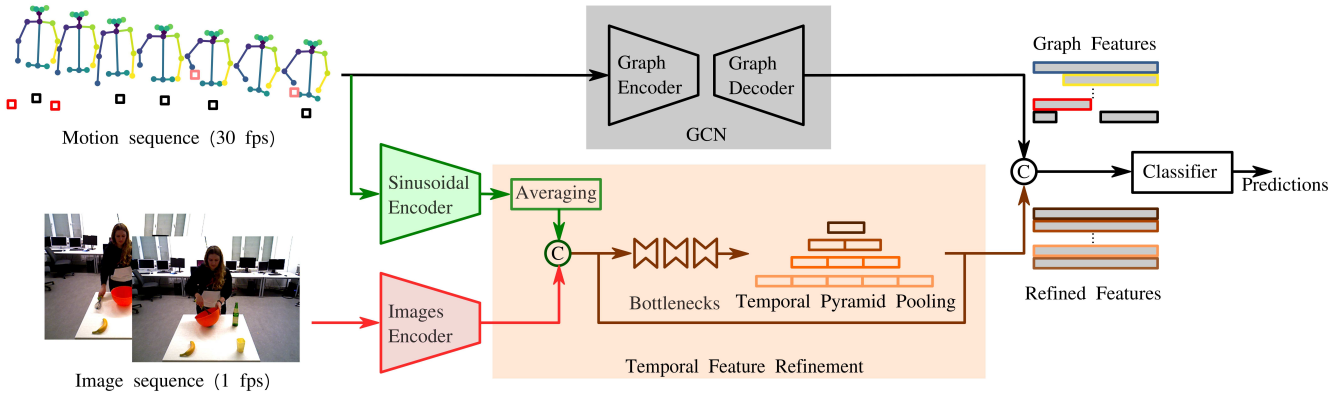


Fig. 1: The framework of Multi-Modal Graph Convolutional Network with motion and image sequences in different temporal resolutions.

segments by half, interpolates the larger segment to match the smaller one to preserve the natural smoothness of human actions, and then combines them into a new sequence. However, its downsampling process disrupts the natural tempo of human actions, leading to inconsistent predictions in speed-sensitive tasks.

To address these limitations, our work introduces SmoothLabelMix, a novel augmentation strategy tailored for action segmentation. Inspired by the observation that human actions transition gradually, SmoothLabelMix smooths the labels along the temporal dimension and blends action segments and their labels along the spatial dimension. This enforces a smooth label distribution that mirrors real-world action dynamics, while simultaneously training the model to handle noisy inputs caused by sensor errors or occlusions.

III. Multi-Modal Graph Convolutional Network

The idea of the Multi-Modal Graph Convolutional Network is jointly optimizing dense spatial features from visual data and dense temporal dynamics from the motion sequences (skeleton and object). Motion inputs (e.g., 30 fps) provide fine-grained motion patterns but suffer from joint tracking noise, while sparse RGB frames (1 fps) offer stable spatial context but lack temporal granularity. Our model bridges these modalities through two core mechanisms: a sinusoidal encoder and temporal feature refinement.

A. Sinusoidal Encoder

To mitigate spatial jitter and enhance compatibility with visual features, we transform the joint position information in the camera coordinate into a continuous sin-cons representation. Given a 3D joint $P_i = (X_i, Y_i, Z_i)$, where $P_i \in \mathbb{R}^{1x3}$, the encoder maps each coordinate into a high-dimensional space using sinusoidal and cosine functions. The encoder first scales these coordinates by a frequency factor determined by the parameters α and β. It then computes sinusoidal sin_e and cosine con_e embeddings for each dimension, resulting in a high-dimensional vector that captures both the position and its spatial context. Mathematically, for each coordinate $c \in \{X_i, Y_i, Z_i\}$, the encoder computes:

$$sin\_e(c) = sin(\frac{\beta \cdot c}{\alpha^{k/d}}), \quad (1)$$

$$cos\_e(c) = cos(\frac{\beta \cdot c}{\alpha^{k/d}}), \quad (2)$$

where k is the index of the embedding dimension, and d is the total number of dimensions. These sinusoidal and cosine embeddings are concatenated to form the final encoded representation P_e:

$$P_{e,i} = concat(sin\_e(P_i), cos\_e(P_i)). \quad (3)$$

B. Temporal Feature Refinement

While position information is encoded using the sinusoidal encoder described earlier, the image information is processed into high-dimensional features through a modified I3D model [23], in which the stride along the time axis is removed because of the limited frames. By concatenating the motion and image features together, a multi-modal feature is formed. The Temporal Feature Refinement module is designed to integrate these multi-modal features into a unified representation and upsample them to the time scale of the graph motion sequence.

As demonstrated in Fig. 1, the module begins by averaging the encoded position features to match the temporal-spatial resolution of the image features. Then the concatenated features are refined by three consecutive bottleneck layers, where the features are compressed in half at the bottleneck place and then remapped back to the high dimensions. The refined features are further upsampled to the motion time scale through a temporal pyramid pooling layer [12] with four parallel temporal pooling blocks. The final refined features are obtained by introducing the interpolated original features into the pooled features.

C. Multi-Modal Graph Convolutional Network

In addition to the motion and image feature refinement stream, the original motion sequence is processed in a graph encoder-decoder network [1]. This network models the structural dependencies between motion elements (skeleton and objects) by representing them as spatial and temporal graphs. The graph nodes are skeleton keypoints and object

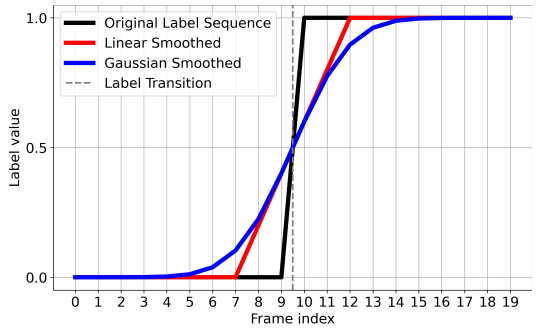


Fig. 2: Comparison of Original, Linear, and Gaussian Smoothed label sequences, all sequences have a label transition at the same frame index.

center points, and the graph edges represent the dynamic relations between nodes. Through multiple layers of graph convolutions and pooling operations, the network learns high-dimensional motion features that preserve both local and global motion contexts. The output of the graph encoder-decoder has the same dimensionality as the motion feature refinement stream, allowing for seamless fusion.

The extracted graph-based motion features and the refined motion-image features are concatenated and passed through a multi-layer classifier. This classification head leverages two cascaded 2D convolutional kernels to complement the strengths of both streams to improve recognition performance. As shown in Figure 1, the complete MMGCN framework integrates the following modules: Sinusoidal Encoder, Image Encoder, Temporal Feature Refinement, Graph Convolutional Network, and a Classifier.

IV. Smooth Label Mix

Human motion smoothly transitions from one action to the next, whereas action labels change in binary form from 0 to 1 or vice versa. To bridge this discrepancy, we propose SmoothLabelMix, a two-stage framework for training sequence models with smoothed labels and intra-batch mixing. The method consists of two key components: (1) **Label Smoothing** using linear or Gaussian filters, and (2) **Weighted Mixing** of input sequences and their corresponding smoothed labels.

The Label Smoothing technique aligns the label transitions with the smooth nature of human motion. Specifically, we implement two types of filters (linear and Gaussian) to process the binary label sequences, as illustrated in Figure 2. These filters create gradual transitions in the labels, better reflecting the continuous nature of human motion. The linear filter applies a uniform averaging over a fixed window, resulting in piecewise-linear transitions, while the Gaussian filter uses a weighted average based on the Gaussian distribution, producing smoother, more natural S-curve transitions. By incorporating these smoothed labels, we aim to improve the alignment between the model's predictions and the inherent continuity of human actions.

The Weighted Mixing technique further enhances model robustness by combining pairs of input sequences and their corresponding smoothed labels within a training batch. Given two input sequences x_1 and x_2, the mixed sequence x_{mix} is computed as:

$$x_{mix} = w \cdot x_1 + (1 - w) \cdot x_2 \tag{4}$$

where w is a weighting factor with $w \sim Beta(\alpha = 0.2)$, balances the contribution of each input. Similarly, the corresponding labels are mixed using the same weighting scheme to maintain consistency between the inputs and annotations. This method enhances the model's ability to generalize across gradual action transitions, reducing abrupt segmentation errors and improving temporal coherence.

V. Experiments and Results

To evaluate the performance of the proposed modules, we conduct the experiments on a public human-objects interaction dataset: Bimanual Actions [24]. On this dataset, we also perform systematic ablation studies to validate four critical design choices:

1) Sinusoidal encoding: ablation with/without temporal positional embeddings.
2) Fusion strategies: early fusion (input-level concatenation), middle fusion (feature-level attention), and late fusion (output-level averaging).
3) Label filters: linear vs. Gaussian filtered label transitions.
4) Weighted mixing: ablation with/without mixing data.

A. Datasets

Bimanual Actions dataset [24] captures fine-grained human-object interactions for kitchen tasks comprising 540 recordings (2 hours 18 minutes total) of bimanual manipulation tasks such as cutting and stirring. The dataset authors propose a leave-one-subject-out cross-validation protocol, where models are evaluated on held-out participants to test generalization to unseen users. For our experiments, we compare against state-of-the-art methods on this benchmark while conducting ablation studies on Subject 1's testset to isolate the impact of our proposed components (e.g., label smoothing, fusion strategies).

For our multi-modal fusion experiments, each input sample integrates a 120-frame motion sequence (comprising 3D body skeletons and object bounding box center points) with 4 RGB images (480×640) uniformly sampled from the video clip. This configuration creates a 30:1 time resolution ratio between the high-frequency motion data (120 frames) and the sparsely sampled visual data (4 frames)

B. Experimental Settings

We evaluate our method using standard metrics to ensure a comprehensive assessment of its performance. For evaluation, we employ metrics such as recognition accuracy, micro-F1, and segmentation accuracy (F1@k). The model is optimized using the SGD optimizer with an initial learning rate of 10^{-4}, which is decayed following a multi-step schedule. Training is conducted on two NVIDIA RTX 2080ti GPUs, leveraging mixed precision to enhance efficiency and reduce memory overhead. We use a batch size of 32 and train for

TABLE I: Ablation studies: the accuracy of framewise prediction and F1@k score of action segmentation using models with different modifications in each unit[a]

Smoothing[b]			Mixing		Refinement		Sinusoidal		Fusion Strategies [c]			Evaluation Metrics (%) [d]				
O	L	G	w	w/o	w	w/o	w	w/o	Early	Mid	Late	Top 1	F1 macro	F1@10	F1@25	F1@50
×				×		×		×				82.91	81.93	91.14	88.39	78.26
	×			×		×		×				83.83	82.82	92.03	89.14	77.30
		×		×		×		×				84.05	83.77	90.63	88.44	77.85
×			×			×		×				84.70	84.68	91.33	89.15	80.61
	×		×			×		×				<u>89.09</u>	<u>89.31</u>	93.41	91.59	80.65
		×	×			×		×				88.92	88.89	<u>94.14</u>	<u>92.52</u>	<u>**84.07**</u>
		×	×			×		×			×	88.48	88.65	92.88	90.95	81.49
		×	×		×			×			×	89.71	89.89	93.80	91.77	83.49
		×	×		×			×		×		89.74	89.95	94.81	92.71	83.64
		×	×		×			×	×			88.42	88.53	92.31	89.82	80.48
		×	×		×		×			×	×	**90.19**	**90.39**	**95.26**	**93.05**	83.87

[a] "Smoothing" means Label Smoothing, "Mixing" represents Weighted Mixing, "Sinusoidal" is the Sinusoidal Encoder, and "Refinement" is the Temporal Feature Refinement block.
[b] "O" is the original label, "L" means the linear filtered label, and "G" means the Gaussian filtered label. "w" is with and "w/o" means without.
[c] Empty entries indicate models trained solely on motion sequences without visual input.
[d] The best results comparing all modifications are in **bold**; The best results among different configurations of label smoothing and weighted mixing are <u>underlined</u>.

60 epochs, applying the SmoothLabelMix data augmentation techniques to improve generalization. The evaluation experiments are conducted on a single GPU.

C. Ablation Study

Our ablation study systematically evaluates the contributions of three core innovations in the framework: SmoothLabelMix data augmentation, the Temporal Refinement module, and the Sinusoidal Encoder within the multi-modal graph convolutional network. The introduction of the image naturally raises the challenge of fusion strategies. Therefore, while analyzing the Sinusoidal Encoder, we also explored the impact of different fusion approaches: early, middle, late, and middle+late. Note that we categorize the implementation of the Sinusoidal Encoder and Temporal Refinement as a middle fusion method, because the Sinusoidal Encoder processes the motion features before fusion within the Temporal Refinement.

The first 6 rows of the ablation study on Table I reveal critical insights into the interplay between label smoothing and data mixing for action recognition and segmentation. Considering the Top1 and micro F1 performance in action recognition, combining label smoothing with data mixing consistently outperforms non-mixed counterparts, with the largest gains observed for linear smoothing + mixing ($89.09/89.31$ vs $83.83/82.82$ without) and Gaussian smoothing + mixing ($88.92/88.89$ vs $84.05/83.77$ without). This demonstrates that mixing enhances model robustness by enforcing consistency between smoothed labels and motion inputs. Considering the F1@k performance, the benefits of Gaussian smoothing + mixing are most pronounced at strict temporal thresholds (F1@50: 84.07 vs. 77.85 without mixing), indicating better handling of action boundaries. Notably, segmentation at loose thresholds (F1@10/25) improves across all configurations, suggesting label smoothing alone helps with coarse temporal alignment, while mixing refines fine-grained transitions. The synergy between Gaussian smoothing and data mixing achieves the best segmentation performance across the configurations of smoothing and mixing techniques, reducing over-segmentation by 5.81% (F1@50) compared to the baseline (the first row). Given its balanced improvement across recognition accuracy and segmentation precision, particularly at challenging temporal thresholds, we adopt the Gaussian label smoothing and data mixing combination for all subsequent experiments.

Beginning at the 7-th row, we introduce the image stream alongside the motion sequence as multi-modal inputs. Rows 7-8 of the ablation study (Table I) evaluate the impact of the proposed Temporal Refinement module, which enhances the integration of visual and motion modalities. Integrating the Temporal Refinement module improves the frame recognition accuracy by 1.23%, while also mitigating over-segmentation, particularly beneficial for F1@50 ($+2.00\%$). The results demonstrate the module's effectiveness in enhancing temporal coherence and segmentation precision. However, the inclusion of image streams still brings disturbance to the model, especially for action segmentation. This observation motivates us to further analyze the configuration of the fusion strategy and the Sinusoidal Encoder.

Rows 8-11 of the ablation study (Table I) reveal critical insights into the efficacy of the fusion strategies and the Sinusoidal Encoder. Early fusion alone performs poorly (Accuracy: 88.42, F1@50: 80.48), as raw skeleton and image features suffer from noise amplification and temporal misalignment. Integrating the Sinusoidal Encoder into a late fusion significantly mitigates these issues, which projects 3D skeleton coordinates into a continuous sin-cos space. The combination of Sinusoidal (middle) and late fusion strategy achieves the best performance (Accuracy: 90.19%, F1@10: 95.26%), demonstrating that sinusoidal encoding stabilizes pose features by smoothing joint trajectories, and late fusion refines boundaries using contextual visual cues. These results underscore the necessity of the Sinusoidal Encoder for the superiority of decoupling early noise suppression from late contextual refinement in multi-modal frameworks.

For the efficiency analysis, our model archives an optimal balance between computational efficiency and performance, operating at 131.0 GFLOPs, significantly lower than using

TABLE II: Comparison of cross-validation results with state-of-the-art methods on the Bimanual Actions dataset [24][a]

Model	Accuracy (%)	F1 macro (%)	F1@10 (%)	F1@25 (%)	F1@50 (%)
Dreher et al. [24]	63.0	64.0	40.6 ± 7.2	34.8 ± 7.1	22.2 ± 5.7
H2O+RGCN [25]	68.0	66.0	–	–	–
Independent BiRNN [26]	74.8	76.7	74.8 ± 7.0	72.0 ± 7.0	61.8 ± 7.3
Relational BiRNN [26]	77.5	80.3	77.7 ± 3.9	75.0 ± 4.2	64.8 ± 5.3
ASSIGN [26]	82.6	79.8	84.0 ± 2.0	81.2 ± 2.0	68.5 ± 3.3
2G-GCN [27]	–	–	85.0 ± 2.2	82.0 ± 2.6	69.2 ± 3.1
PGCN [12]	86.8	83.9	$88.5 \pm \mathbf{1.1}$	85.5 ± 2.0	77.0 ± 3.4
TFGCN [1]	88.4	88.6	93.7 ± 1.2	$91.9 \pm \mathbf{1.6}$	$85.4 \pm \mathbf{2.9}$
MMGCN (ours)	**89.3**	**89.3**	$\mathbf{94.5} \pm 1.6$	$\mathbf{92.8} \pm 1.9$	$\mathbf{86.1} \pm 3.1$

[a] The models are cross validated on the leave-one-subject-out benchmark, the best results of each class are in **bold**. The Accuracy and F1 micro results are averaged, and F1@k are listed with mean and standard deviation. A smaller standard deviation value indicates greater robustness in the model.

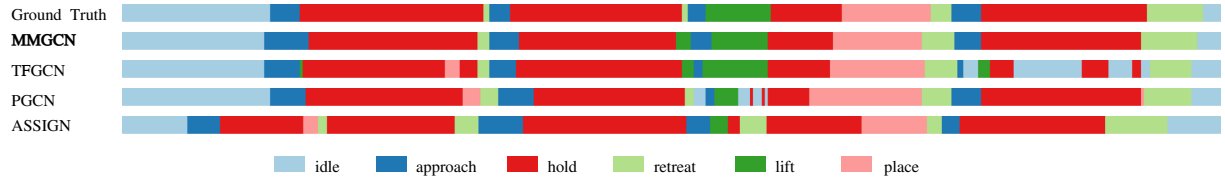
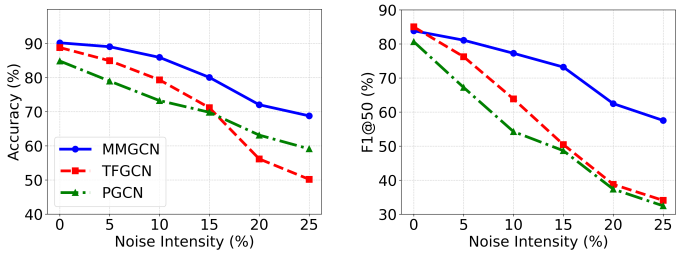


Fig. 3: Qualitative comparison of action segmentation results: Visualized segmentation outputs of different models for the breakfast cereal preparation scenario [24], illustrating the temporal alignment and accuracy of predicted action boundaries.



(a) (b)

Fig. 4: Comparison of the robustness of three models under increasing noise intensities. The figures illustrate how (a) action recognition accuracy and (b) segmentation F1@50 vary as noise levels rise.

higher visual temporal resolution. For instance: processing motion and visual modalities at a resolution ratio of 30 : 2 increases FLOPs to 220.0 G, while uniform high-resolution processing (30 : 30) escalates costs to 2687.3 GFLOPs, a $20.5\times$ increase. This exponential scaling highlights the inefficiency of naively aligning modalities at high frequencies. While our framework introduces moderate computational overhead compared to models using pure motion sequences, it retains real-time performance (81.0 FPS on a single GPU) by strategically decoupling temporal resolutions.

D. Comparison with states-of-the-art

The proposed MMGCN is compared with state-of-the-art methods in the field of human action segmentation on the Bimanual Actions dataset [24]. The compared methods include the method proposed by Dreher *et al.* [24], H2O+RGCN [25], BiRNN [26], 2G-GCN [27], PGCN [12] and TFGCN [1].

The cross-validation results are listed in Table II. Our proposed MMGCN achieves state-of-the-art performance across all evaluation metrics, as validated on the leave-one-subject-out benchmark. Compared to TFGCN [1], the previous best method, MMGCN improves accuracy from 88.4% to 89.2% and F1 macro from 88.6% to 89.1%, with significant gains in boundary-sensitive metrics: F1@10 increases from 93.7 to 94.3, F1@25 from 91.9 to 92.7, and F1@50 from 85.4 to 85.9. Compared to PGCN [12], which achieved 88.5%, 85.5%, and 77.0% action segmentation performance, MMGCN demonstrates substantial improvements, particularly at higher overlap thresholds, indicating better temporal localization of action boundaries. These improvements highlight MMGCN's robustness to temporal misalignment and the efficacy of the multi-modal fusion strategy. Despite achieving superior performance, MMGCN exhibits a slightly higher standard deviation compared to TFGCN [1], particularly at the F1@10 and F1@25. This increased variability might be attributed to differences in model architecture and sensitivity to variations in motion and image inputs. This observation motivates further experiments under noise-induced conditions to evaluate the model's robustness, ensuring consistent performance across diverse scenarios.

To evaluate the robustness of MMGCN under noisy conditions, we selected the most common noise - misdetection. Specifically, on the testset of subject one of the Bimanual Actions dataset [24], the motion nodes are randomly removed per frame. The removal rate (noise intensity) increases from 0% to 25%. The results of robustness are illustrated in Fig. 4. As noise intensity increases, the gap in action (a) recognition and (b) segmentation accuracy between the MMGCN and the other two models progressively widens. Notably, the MMGCN consistently maintains the highest performance across all noise levels and, in the segmentation case, overtakes competing models to secure the top position. This trend highlights the superior resilience of our approach to partially missing inputs.

Figure 3 presents the qualitative comparison of action segmentation results for the breakfast cereal preparation scenario from the Bimanual Actions dataset [24]. Among

the compared models, our proposed MMGCN produces predictions that most closely align with the ground truth. It successfully mitigates the over-segmentation issues, which are commonly observed in other methods. This demonstrates the model's effectiveness in maintaining coherent action segments. However, MMGCN overlooks actions with short durations or extends their intervals, leading to temporal shifts in action boundaries.

VI. CONCLUSIONS

In this study, we propose the Multi-Modal Graph Convolutional Network (MMGCN), a novel framework for robust human action segmentation that harmonizes high-frequency motion data (e.g., 30 fps) with low-frequency visual cues (e.g., 1 fps) via a Sinusoidal Encoder and a mid-stage fusion strategy. The Sinusoidal Encoder projects 3D joint coordinates into a continuous spatial-temporal embedding, while our hybrid fusion architecture integrates motion and visual features at divergent temporal resolutions, preserving fine-grained dependencies and computation efficiency. Complemented by SmoothLabelMix, a data augmentation technique that synthesizes realistic motion-noise transitions, MMGCN achieves state-of-the-art performance across the action recognition and segmentation benchmark.

The model's sensitivity to variations in input data motivates further research into enhancing its robustness. Future work will focus on investigating the model's performance under noisy and incomplete input conditions by incorporating noise injection techniques during training. Additionally, we plan to explore alternative fusion mechanisms, advanced attention models, and extended benchmarks on diverse datasets to validate the model's generalizability and practical applicability across different action recognition tasks.

ACKNOWLEDGMENT

We extend our sincere gratitude to Timothy Ju Kin, Ng for his invaluable technical insights and support.

REFERENCES

[1] H. Xing and D. Burschka, "Understanding human activity with uncertainty measure for novelty in graph convolutional networks," *The International Journal of Robotics Research*, p. 02783649241287800, 2024.

[2] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.

[3] H. Tani, "Graph convolutional networks with minimal appearance information for action recognition," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 388–394.

[4] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High five: Recognising human interactions in tv shows." in *BMVC*, vol. 1, no. 2, 2010, p. 33.

[5] C. Xu, C. Xiong, and J. J. Corso, "Streaming hierarchical video segmentation," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 2012, pp. 626–639.

[6] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[7] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3575–3584.

[8] Y. Ishikawa, S. Kasai, Y. Aoki, and H. Kataoka, "Alleviating over-segmentation errors by detecting action boundaries," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2322–2331.

[9] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[10] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 026–12 035.

[11] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[12] H. Xing and D. Burschka, "Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5195–5201.

[13] J. Luo, W. Wang, and H. Qi, "Spatio-temporal feature extraction and representation for rgb-d human action recognition," *Pattern Recognition Letters*, vol. 50, pp. 139–148, 2014.

[14] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2123–2129, 2015.

[15] S. Kim, D. Ahn, and B. C. Ko, "Cross-modal learning with 3d deformable attention for action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10 265–10 275.

[16] J. Wang, L. Xia, and X. Wen, "Cmf-transformer: cross-modal fusion transformer for human action recognition," *Machine Vision and Applications*, vol. 35, no. 5, p. 114, 2024.

[17] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[20] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[21] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim, "Videomix: Rethinking data augmentation for video classification," *arXiv preprint arXiv:2012.03457*, 2020.

[22] L. Xiang and Z. Wang, "Joint mixing data augmentation for skeleton-based action recognition," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[23] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[24] C. R. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 187–194, 2019.

[25] D. Lagamtzis, F. Schmidt, J. Seyler, T. Dang, and S. Schober, "Exploiting spatio-temporal human-object relations using graph neural networks for human action recognition and 3d motion forecasting," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7832–7838.

[26] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 041–16 050.

[27] T. Qiao, Q. Men, F. W. Li, Y. Kubotani, S. Morishima, and H. P. Shum, "Geometric features informed multi-person human-object interaction recognition in videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 474–491.