Multimodal Feature Collaboration and Fusion for Fine-grained Action Recognition

Xinyu Bian[1], Dongliang Chang[2], Yuqi Yang[1], Lei Chen[2], and Zhanyu Ma[1]

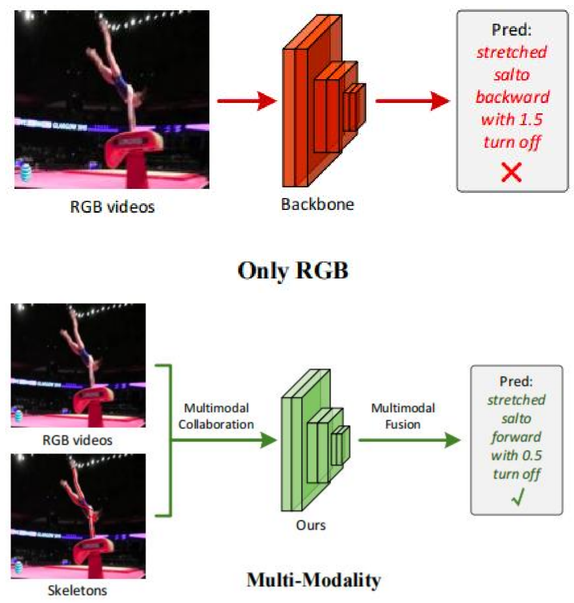
[1] Beijing University of Posts and Telecommunications
[2] Department of Automation Tsinghua University

BMVC 2025

Introduction

Fine-grained action recognition remains a challenging task due to subtle inter-class differences and significant intra-class variations. Existing methods often rely on single-modality inputs such as RGB or skeleton data, which limits feature completeness and robustness. To overcome these limitations, we propose a Multimodal Feature Collaboration and Fusion framework that enhances cross-modal complementarity for efficient fine-grained action understanding.

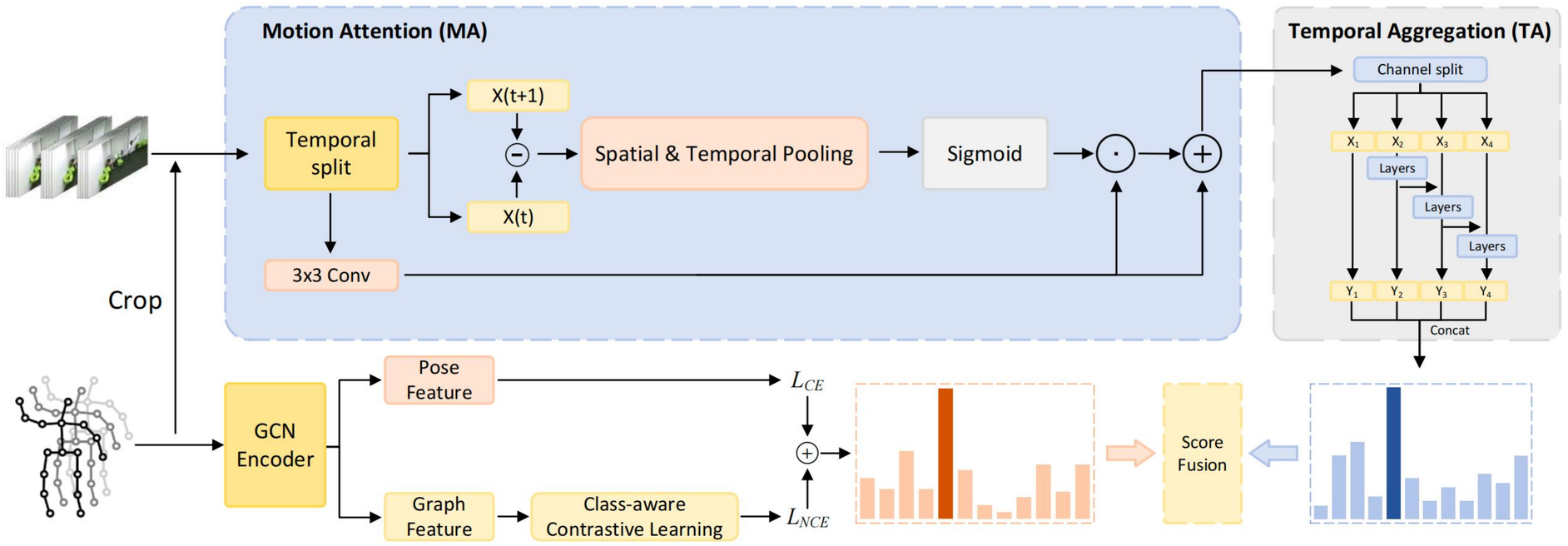
Our framework leverages skeleton-guided cropping to suppress redundant background in RGB videos, integrates a Spatio-Temporal Attention and Aggregation module to capture both short-term motion and long-term temporal dependencies, and introduces a Class-aware Contrastive Learning scheme to strengthen feature discriminability across similar actions. Through score-weighted fusion of multimodal features, our approach achieves state-of-the-art performance with reduced computational cost.



Method

Spatio-Temporal Attention and Aggregation (STAA)

A lightweight dual-branch structure captures fine-grained motion cues. The Motion Attention submodule enhances dynamic local patterns via inter-frame differencing and adaptive weighting, while the Temporal Aggregation submodule extends temporal receptive fields through hierarchical residual convolutions for global reasoning.



Multimodal Feature Fusion Strategy

- We first perform skeleton-guided cropping on RGB frames, where skeleton joint coordinates localize motion regions and remove background redundancy. This operation achieves data-level fusion, allowing skeleton information to guide RGB feature extraction from the very beginning.
- At the decision level, a score-weighted fusion combines recognition results from RGB and skeleton modalities according to their confidence, ensuring that each modality contributes optimally to the final prediction.

Experiments

Modal representation

- R: RGB modal
- S: Skeleton modal
- R+S: RGB and Skeleton modals

Computation complexity

Methods	FLOPs
VPN	107.9G
VPN++	125.8G
TSMF	85.4G
MMNet	89.2G
RGBPose-C3D	1410G
Ours	**35.1G**

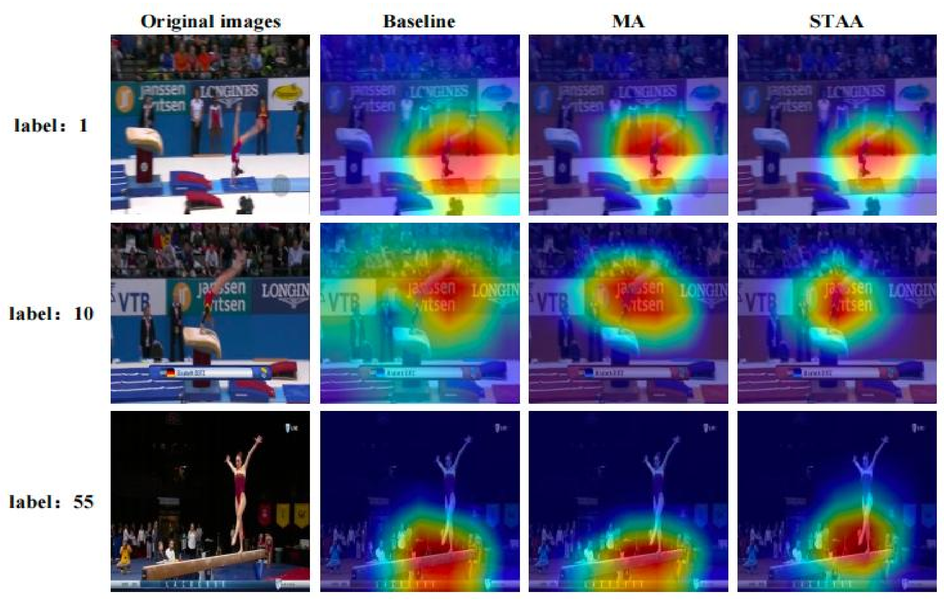
Ablation Study

Methods	X-Sub
Baseline	92.3
Pose	92.8
Baseline + Crop	95.3
Baseline + Crop + MA	96.9
Baseline + Crop + STAA	97.6
Ours	**98.8**

Comparison Results on FineGym

Methods	Modality	FineGym
I3D	R	75.6
TSM	R	88.4
MTN	R	91.8
MS-G3D	S	92.0
InfoGCN	S	92.0
PoseC3D	S	93.2
Ske2Grid	S	91.8
RGBPose-C3D	R+S	95.6
Ours	R+S	**96.0**

Visualization Results on FineGym



Comparison Results on NTU60

Methods	Modality	NTU60 X-Sub	NTU60 X-View
Glimpse Clouds	R	86.6	93.2
Action Machine	R	94.3	97.2
TimeSformer	R	93.0	97.2
ST-GCN	S	81.5	88.3
2s-AGCN	S	88.5	95.1
Shift-GCN	S	90.7	96.5
MS-G3D	S	91.5	96.2
CTR-GCN	S	92.4	96.8
InfoGCN	S	92.7	96.9
FR-Head	S	92.8	96.8
Ske2Grid	S	90.1	95.1
DSTSA-GCN	S	92.8	97.0
S-Res-LSTM	R+S	90.0	96.3
VPN	R+S	93.5	96.2
MMNet	R+S	96.0	98.8
RGBPose-C3D	R+S	95.3	96.4
STAR++	R+S	92.7	97.6
VT-BPAN	R+S	95.4	97.7
π-ViT	R+S	96.3	99.0
EPAM-Net	R+S	96.5	99.0
Ours	R+S	**98.8**	**99.7**