Multimodal Feature Collaboration and Fusion for Fine-grained Action Recognition

Xinyu Bian[1]
bianxinyu@bupt.edu.cn
Dongliang Chang[1*]
changdongliang@pris-cv.cn
Yuqi Yang[1]
yangyuqi@bupt.edu.cn
Lei Chen[2]
leichenthu@tsinghua.edu.cn
Zhanyu Ma[1]
mazhanyu@bupt.edu.cn

[1] School of Artificial Intelligence
Beijing University of Posts and Telecommunications
Beijing, China

[2] Department of Automation
Tsinghua University
Beijing, China

Abstract

Fine-grained action recognition faces challenges of subtle inter-class differences and significant intra-class variations in complex real-world scenarios. Existing methods are often limited by insufficient feature representation capabilities of single modalities or redundant computational costs. In this paper, we propose a Multimodal Feature Collaboration and Fusion framework to achieve efficient recognition through cross-modal guided spatial and temporal modeling. For the RGB modality, skeleton joint coordinates collaboratively localize and crop action regions to suppress background noise, and design a lightweight Spatio-Temporal Attention and Aggregation (STAA) that combines short-term motion excitation and long-term temporal modeling to enhance feature extraction efficiency and accuracy. For the skeleton modality, we propose a Class-aware contrastive learning that dynamically selects high-similarity samples to optimize feature discriminative boundaries. Finally, a score-weighted fusion strategy integrates the optimized representations from both modalities. Experiments on the FineGym and NTU RGB+D benchmarks demonstrate impressive performance improvements over state-of-the-art methods.

1 Introduction

Fine-grained human action recognition is a core field of computer vision with intelligent surveillance, sports analysis, and human-computer interaction [27, 46, 50, 56]. With the increasing requirement for understanding specific actions in real-world applications, fine-grained action recognition has gradually become a research focus. The researchers have proposed a variety of self-supervised and unsupervised methods [24, 39, 54, 58, 61] to mine semantic information inherent in the data. However, distinguishing visually or kinematically similar actions is critical but challenging, due to subtle inter-class differences and large

*Corresponding author

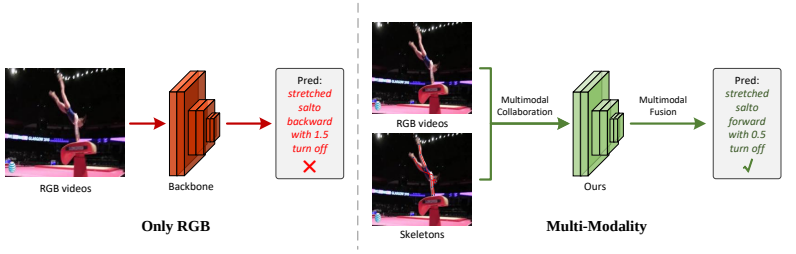

Figure 1: Comparison of a single RGB modality and our multimodal collaboration and fusion framework integrating RGB and skeletaon data.

intra-class variability. This subtle differentiation is critical for many practical applications. Similar to fine-grained visual classification [8, 18, 38, 55], where categories are highly similar and distinguished by subtle and localized cues, fine-grained action recognition faces analogous challenges. Reliance on single-modality data, such as RGB videos or skeleton data alone, often limits recognition performance. RGB videos offer rich context but suffer from background noise, while skeleton data capture accurate joint kinematics but omit object interactions.

To address these limitations, researchers have adopted the multimodal strategy to exploit the strengths of RGB and skeleton data [1, 6, 32]. These two modalities exhibit complementarity. However, previous methods typically employ concatenation-based network components to facilitate feature fusion [28], overlooking data-level complementarity. Moreover, existing methods typically employ single-modality networks for multimodal modeling without implementing targeted optimization for each modality-specific branch, leading to suboptimal performance and computational redundancy.

To overcome these challenges, this paper propose a Multimodal Feature Collaboration and Fusion approach in which we optimize feature extraction for both RGB and skeleton modalities and enhance fine-grained action recognition through collaborative and fusion strategies. For the RGB modality, we employ skeleton-guided video region cropping to achieve multimodal collaboration, using joint coordinates to locate and crop key motion regions and suppress background noise. We then introduce a Spatio-Temporal Attention and Aggregation (STAA) that concurrently extracts short-term and long-term features, thereby capturing comprehensive motion dynamics. For the skeletal modality, we adopt a class-aware contrastive learning that dynamically selects samples from similar categories for training, strengthening the model's ability to discriminate among analogous actions. Building on these optimized representations, we apply a weighted score fusion strategy, assigning weights according to each modality's recognition confidence to integrate complementary information. Moreover, to reduce computational cost, we compress the RGB videos into a smaller sequence of frames prior to feeding it into the backbone network, minimizing input data while maintaining an optimal balance between recognition performance and efficiency.

The contributions of this paper can be summarized as follows:

1. We propose a lightweight spatio-temporal feature extraction approach (STAA), where the short-term module focuses on motion regions while the long-term module aggregates temporal features, effectively expanding the receptive field to capture richer features.

2. We present a multimodal fusion method that integrates RGB and skeleton features to exploit their complementary information. We leverage STAA to enhance spatio-temporal feature extraction and employ contrastive learning to bolster the model's discriminative power. Finally, a weighted fusion strategy is used to achieve the optimal score.

3. By integrating the above methods with skeleton-guided RGB cropping, we construct our entire network framework, achieving competitive performance on two large action recognition datasets.

2 Related Works

2.1 Video Action Recognition

With the remarkable success of deep learning methods [16, 20, 22, 29, 36] in RGB-based recognition tasks, researchers have begun exploring the application of deep networks to video action recognition. Simonyan *et al*. [46] uses parallel spatial RGB flow and temporal optical flow streams, fusing their scores to significantly boost action recognition accuracy. Feichtenhofer *et al*. [22] employ cross-stream multiplicative connections at multiple convolutional layers to merge RGB and flow features, enabling mutual reinforcement of shallow and deep representations. Diba *et al*. [16] extend all convolutions and pooling into the spatio-temporal domain and introduce a Temporal Transition Layer (TTL) that learns variable-depth temporal kernels to flexibly capture motions of different durations. While deep architectures can capture complex spatio-temporal patterns, they often come with heavy computational burden. To alleviate this, Temporal Segment Network (TSN) [51] divides a video into a fixed number of segments, sparsely samples one or a few frames per segment, and applies a lightweight CNN to extract local temporal representations, dramatically reducing inference cost. In contrast, our method achieves higher recognition accuracy with even lower computational overhead.

2.2 Skeleton Action Recognition

Skeleton data contains critical information for action recognition. Common methods [10, 12, 35, 45, 57, 63] employ Graph Convolutional Networks (GCN) to model spatio-temporal dependencies. Yan *et al*. [57] propose a spatial-temporal graph convolutional network(ST-GCN) that builds fixed skeleton graphs and applies GCN over spatio-temporal dimensions. Shi *et al*. [45] design a two-stream adaptive GCN that models joint and bone information, using data-driven edge importance weighting to better capture motion patterns. Chen *et al*. [10] propose CTR-GCN that introduces dynamic channel-wise skeleton topology learning, enabling fine-grained joint dependency reasoning for each channel to surpass previous methods. To enrich pose details, cross-form attention [53] and hierarchical edge decomposition [30] have been adopted. For distinguishing similar actions, InfoGCN [12] leverages an information bottleneck principle to derive compact yet semantically rich representations, and ActCLR [34] uses actionlet-based contrastive learning.

2.3 Multimodal fusion methods

Multimodal approaches [6, 49, 59, 65] enhance action recognition accuracy and robustness by integrating information from diverse sources. Existing fusion strategies [9, 17, 19, 60] typically integrate multimodal data during the feature extraction process. Zhu *et al*. [65] propose a two-stage fusion framework combining skeleton sequences and RGB frames to capture complementary features for action recognition. The Model-based Multimodal Network

(MMNet) [6] employs a model-driven framework to integrate RGB and skeleton modalities, significantly enhancing action recognition accuracy. Most existing methods focus on feature-level information fusion while neglecting the guiding role of skeleton data at the data level for video analysis. Our method integrates RGB and skeleton modalities at the input stage by leveraging skeleton information to localize critical action regions in RGB videos. The framework further enhances recognition performance by combining modality outputs using a score fusion scheme.

3 Methodology

Inspired by the Temporal Shift Module (TSM) [33], our method first divides the input RGB videos into s equal segments and randomly samples several consecutive frames from each segment. In Section 3.1, we introduce the skeleton-guided cropping strategy, which reduces background noise by cropping all sampled frames. In Section 3.2, we present the Spatio-Temporal Attention and Aggregation (STAA) module, which achieves lightweight extraction of fine-grained action features through short-term motion attention and long-term temporal aggregation. In Section 3.3, we provide a detailed account of the Class-aware Contrastive Learning module, in which the GCN is employed to extract skeleton features and the class-aware sample strategy selects highly similar negative samples to strengthen the model's discriminative power. In Section 3.4, we merge the RGB and skeleton scores using a weighted fusion approach to produce the final action recognition scores. The framework is shown in Figure 2.

3.1 Skeleton-Guided Cropping

To mitigate the interference of background clutter on motion feature extraction, we apply skeleton-guided cropping. Specifically, let the coordinates of skeleton joints be denoted as $x \in X$ and $y \in Y$. We extract an activity region of width $w = \max(X) - \min(X)$ and height $h = \max(Y) - \min(Y)$. To preserve action completeness while suppressing background, we expand both the height and width of the bounding box by α along its dimensions, then clamp its coordinates to the image borders. The cropping box is uniformly applied across all RGB frames to highlight motion regions and maintain background consistency.The final cropped motion region X_{region} and Y_{region} can be expressed as:

$$X_{region} = [\max(\min(X) - \alpha w, 1), \min(\max(X) + \alpha w, W)], \tag{1}$$

$$Y_{region} = [\max(\min(Y) - \alpha h, 1), \min(\max(Y) + \alpha h, H)], \tag{2}$$

where W and H denote the width and height of the original RGB frames.

3.2 Spatio-Temporal Attention and Aggregation (STAA)

The Spatio-Temporal Attention and Aggregation (STAA) module cascades two core components: Motion Attention (MA) and Temporal Aggregation (TA). The MA module enhances localized motion patterns through feature differencing and dynamic channel attention, while the TA module captures global temporal dependencies via a grouped residual convolutional chain.

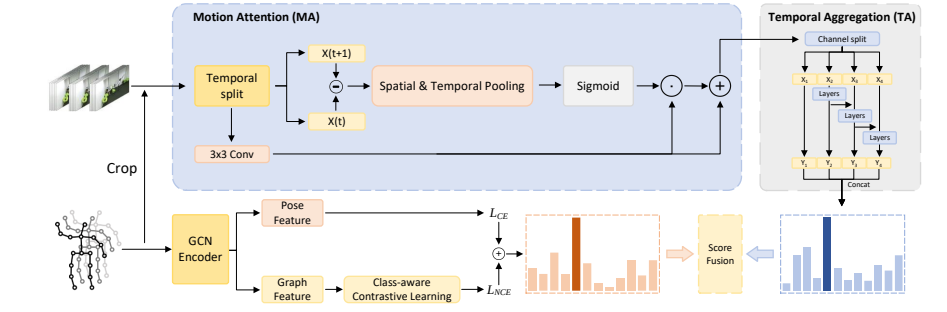

Figure 2: The framework of our method. The cropped frames are fed into MA and TA for action prediction. Skeleton datas are input into the GCN for feature extraction, and class-aware contrastive learning is utilized for classification prediction. Finally, the scores from both RGB and skeleton datas are merged with fixed weights to produce the multimodal score.

3.2.1 Motion Attention (MA)

The MA module enhances short-term motion information via spatio-temporal modeling. It extracts motion-sensitive channels via integrates inter-frame difference into spatio-temporal feature learning. Specifically, the input feature M has a shape of $[N,T,C,H,W]$, where N denotes the batch size, T represents the dimension, C indicates the number of feature channels, while H and W correspond to the spatial height and width respectively. We first compute inter-frame differences to extract motion features, the motion feature at time t is computed as:

$$\mathcal{F}(t) = M(t+1) - M(t), \quad 1 \leq t \leq T-1, \tag{3}$$

where $\mathcal{F}(t) \in \mathbb{R}^{N\times C\times H\times W}$ represents the motion feature at time t. Subsequently, the final motion matrix $\mathcal{F}$ is constructed by concatenating all motion features $[\mathcal{F}(1),...,\mathcal{F}(T)]$. The features are processed through temporal and spatial pooling to condense motion information and generate attention masks, as shown in Equation 4.

$$\mathcal{F}_{mask} = \sigma\big(\text{TemporalPool}(\text{SpatialPool}(\mathcal{F}))\big), \tag{4}$$

where $\mathcal{F}_{mask}$ denotes the attention mask, both SpatialPool and TemporalPool correspond to maxpooling operations, σ is the sigmoid function, and $\odot$ indicates element-wise multiplication. Finally, the original features are adaptively weighted through residual connections, enabling the model to selectively emphasize dynamic motion information while simultaneously preserving the semantic integrity of static scenes:

$$M_{map} = M \odot \mathcal{F}_{mask} + \phi(M), \tag{5}$$

where M_{map} is the motion map, $\phi(\cdot)$ is the original features processed by a 3×3 convolution and $\odot$ represents element-wise multiplication. This architecture effectively reduces computational overhead through spatio-temporal modeling while preventing excessive suppression of background context via the residual mechanism.

3.2.2 Temporal Aggregation (TA)

Traditional methods rely on stacked local temporal convolutions to indirectly model long-range dependencies, but information between distant frames tends to be weakened in deep

networks. To address this, we propose a Temporal Aggregation (TA) module inspired by Res2Net [23]. Given an input motion map M_{map}, we evenly split its channels into four fragments $[v_1, v_2, v_3, v_4]$. The last three fragments are sequentially processed with 1D depthwise separable and 2D spatial convolutions. This residual design expands the temporal receptive field through hierarchical fusion, with earlier outputs progressively fed into later computations as:

$$[v_1, v_2, v_3, v_4] = \text{ChannelSplit}(M_{map}), \tag{6}$$

$$\begin{aligned} q_1 &= v_1, & q_2 &= \text{Conv}_{\text{dep}}(\text{Conv}_{\text{sc}}(v_2)), \\ q_3 &= \text{Conv}_{\text{dep}}(\text{Conv}_{\text{sc}}(v_3 + q_2)), & q_4 &= \text{Conv}_{\text{dep}}(\text{Conv}_{\text{sc}}(v_4 + q_3)), \end{aligned} \tag{7}$$

$$Q = \text{Concat}(q_1, q_2, q_3, q_4), \tag{8}$$

where Q is the aggregated feature, q_1, q_2, q_3, q_4 are the output of four fragments, ChannelSplit layer denotes a 1×1 convolution, Conv_{sc} denotes a spatial convolution with 3×3 kernels, Conv_{dep} refers to a depthwise separable convolution, and $\text{Concat}(\cdot)$ concatenates its inputs along the channel axis. TA enhances long-term temporal modeling through a group-wise residual design, enabling efficient global reasoning for complex actions.

3.3 Class-aware Contrastive Learning

We adopt CTR-GCN [10] as the GCN encoder. Following the method [26], our method integrates the graph contrastive learning into both classification and contrastive phases, yielding pose and graph features. We extract the graph features G as described in Equation 9:

$$G = g(f(\sum_{k=1}^{K_s} H W_k A_k)), \tag{9}$$

where K_s represents the number of learnable topological subgraphs, A_k is the adjacency matrix of the k-th subgraph, H is the input skeleton sequence, and W_k is the corresponding weight matrix. The function f represents the GCN encoder, and g denotes the graph projection head.

3.3.1 Sampling Strategy

To address the challenge of minimal inter-class differences, we introduce the class-aware sampling strategy to select similar negative samples. Similar to [25], we construct a memory bank with a first-in-first-out update strategy to store negative sample features, allowing comparisons across a broader feature space. We calculate pairwise similarity scores and the mean similarity score h_c for each category, as follows:

$$h_c = \frac{1}{M_c} \sum_{m=1}^{M_c} \left(\frac{\phi(z_j) \cdot \phi(z_m)}{\|\phi(z_j)\| \|\phi(z_m)\|} \right), \tag{10}$$

where $\phi(\cdot)$ represents the feature representation of sample, z_j and z_m are feature vectors of the anchor sample and the m-th negative sample from the class c, $\cdot$ is the dot product operation, $\|\cdot\|$ is the Euclidean norm, and M_c is the total number of negative samples in class c.

Based on mean similarity scores, categories are ranked, and the top ones are selected to sequentially choose N_c negative samples. The strategy prioritizes comparisons between highly similar samples, encouraging tighter positive clustering and better negative separation.

3.3.2 Contrastive Learning Loss

We employ the cross-entropy loss and the InfoNCE loss [40] to jointly optimize the contrastive learning process. Cross-entropy loss $\mathcal{L}_{\text{CE}}$ is used to supervise classification predictions, and InfoNCE loss $\mathcal{L}_{\text{NCE}}$ is used to bring positive pairs closer while pushing negative pairs away, as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{i} p_i \log \hat{p}_i, \tag{11}$$

where p_i is the true label of the i-th class, and $\hat{p}_i$ denotes the predicted probability.

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\text{sim}(z, z_+)/\tau)}{\exp(\text{sim}(z, z_+)/\tau) + \sum^{N_c} \exp(\text{sim}(z, z_-)/\tau)}, \tag{12}$$

where z is the anchor sample, z_+ corresponds to the positive samples sharing the same label as z, z_- refers to the negative samples with different labels, N_c is the total number of negative samples, τ is the temperature parameter, $\text{sim}(\cdot,\cdot)$ is the cosine similarity. The overall loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{NCE}}, \tag{13}$$

3.4 Score fusion

We fuse the prediction scores from RGB and skeleton modalities by computing a weighted sum of per-modality class probability scores and selecting the class with the maximum aggregate score:

$$\mathcal{K} = \arg\max \sum_{n=1}^{N_s} \lambda_n \mathcal{S}_n, \tag{14}$$

where $\mathcal{K}$ is the predicted action class, N_s is the number of subnetworks, $\mathcal{S}_n$ is the class probability score produced by the n-th network, and λ_n is the weight corresponding to the subnetwork. The weighted score fusion effectively combines RGB and skeleton features, significantly enhancing action recognition accuracy.

4 Experiments

4.1 Datasets

The proposed method is evaluated on both the large-scale action recognition dataset NTU RGB+D and the fine-grained dataset FineGym.

NTU RGB+D 60 (NTU60) [43] is a large-scale action recognition dataset containing 56,880 action sequences collected from 60 action classes. Two standard evaluation protocols are typically adopted: (1) Cross-Subject (X-Sub): Training data comes from 20 subjects, while testing data uses the remaining 20 subjects. (2) Cross-View (X-View): Training data is captured by cameras 2 and 3, with testing data from camera 1.

FineGym [44] is a fine-grained action recognition dataset comprising 29,000 videos across 99 action categories. It collects 10 different events in gymnastics and defines 15 groups, further specifying 530 distinct elements.

Table 1: Comparison of classification accuracy with state-of-the-art methods on the NTU60 dataset. Notations: R: RGB, S: Skeleton.

Methods	Modality	NTU60 X-Sub	NTU60 X-View
Glimpse Clouds[3]	R	86.6	93.2
Action Machine[24]	R	94.3	97.2
TimeSformer[4]	R	93.0	97.2
ST-GCN[27]	S	81.5	88.3
2s-AGCN[25]	S	88.5	95.1
Shift-GCN[11]	S	90.7	96.5
MS-G3D[32]	S	91.5	96.2
CTR-GCN[10]	S	92.4	96.8
InfoGCN[12]	S	92.7	96.9
FR-Head[62]	S	92.8	96.8
Ske2Grid[7]	S	90.1	95.1
DSTSA-GCN[13]	S	92.8	97.0
S-Res-LSTM[47]	R + S	90.0	96.3
VPN[14]	R + S	93.5	96.2
MMNet[6]	R + S	96.0	98.8
RGBPose-C3D[21]	R + S	95.3	96.4
STAR++[2]	R + S	92.7	97.6
VT-BPAN[48]	R + S	95.4	97.7
π-ViT[41]	R + S	96.3	99.0
EPAM-Net[1]	R + S	96.5	99.0
Ours	R + S	**98.8**	**99.7**

4.2 Implementation details

For the RGB modality, we adopt 8 segments, sampling 6 frames from each segment. The cropped RGB frames are resized to a resolution of 256 $\times$ 256 pixels. α is set to 0.3. The backbone network utilizes a ResNet50-based TSM pretrained on ImageNet, optimized using SGD with an initial learning rate of 0.01 and momentum of 0.9. Training on both the FineGym and NTU60 datasets is conducted for 60 epochs. For the skeleton modality, GCN encoder employs CTR-GCN architecture, while other hyperparameters remain consistent with the RGB modality. The temperature hyperparameter τ is set to 0.8. In the contrastive learning, we set the number of positive samples N_+ to 128 and negative samples N_c to 512. All experiments are implemented on RTX 3090 GPUs using the PyTorch framework.

4.3 Comparisons with the state-of-the-art methods

In Table 1, our method achieves 98.8% accuracy on X-sub and 99.7% on X-view, which is 2.3% on X-Sub and 0.7% on X-View above the previous best, outperforming existing RGB and skeleton fusion methods. In Table 2, our method reaches 96.0% accuracy on FineGym dataset, which is 7.6% above the baseline and 0.4% above the previous best. These results demonstrate that multimodal feature collaboration and fusion effectively improve fine-grained action recognition and enhance the model's ability to distinguish subtle action differences.

As shown in Table 3, we compare the computational complexity of our method against

Table 2: Comparison of classification accuracy with state-of-the-art methods on FineGym dataset. Notations: R: RGB, S: Skeleton.

Methods	Modality	FineGym
I3D[32]	R	75.6
TSM[33]	R	88.4
MTN[51]	R	91.8
MS-G3D[37]	S	92.0
InfoGCN[12]	S	92.0
PoseC3D[21]	S	93.2
Ske2Grid[7]	S	91.8
RGBPose-C3D[21]	R + S	95.6
Ours	R + S	**96.0**

Table 3: Comparisons of computation complexity with the state-of-the-art methods.

Methods	FLOPs
VPN[14]	107.9G
VPN++[15]	125.8G
TSMF[5]	85.4G
MMNet[6]	89.2G
RGBPose-C3D[21]	1410G
Ours	35.1G

Table 4: Comparison results of the proposed components. Notations: R: RGB, S: Skeleton.

Methods	Modality	NTU60 X-Sub	NTU60 X-View
Baseline[33]	R	92.3	95.8
Pose	S	92.8	96.7
Baseline + Crop	R + S	95.3	98.3
Baseline + Crop + MA	R + S	96.9	98.9
Baseline + Crop + STAA	R + S	97.6	99.1
Ours	R + S	**98.8**	**99.7**

the state-of-the-art methods. The results demonstrate that our method delivers superior accuracy with reduced computational cost.

Overall, the comparison against RGB, skeleton, and fusion methods shows our method achieves better performance. On the FineGym dataset, although its accuracy matches that of RGBPose-C3D, RGBPose-C3D incurs higher computational cost. In contrast, our method reduces computation while improving accuracy. We also confirm the model's advantage on the coarse-grained NTU60 dataset, further extending the applicability of our method.

4.4 Ablation studies

We perform ablation experiments to evaluate the effectiveness of the components using TSM [33] as the baseline. For the RGB modality, we assess the impact of STAA, MA, and TA modules, while for the skeleton modality, we compare the performance of different negative sampling strategies.

4.4.1 Effectiveness of Proposed Main Components

In Table 4, we demonstrate the effectiveness of our proposed components on NTU60 dataset. First, the baseline using only RGB videos achieves 92.3% and 95.8% on X-sub and X-view. Adding skeleton modality to guide motion region cropping improves recognition accuracy by 3.0% on X-sub and 2.5% on X-view. Further integrating MA boosts X-sub accuracy by 1.6%. Subsequently, the addition of STAA improves the X-view accuracy by 0.8%. Finally, our method achieves improvements of 6.5% on X-sub and 3.9% on X-view.

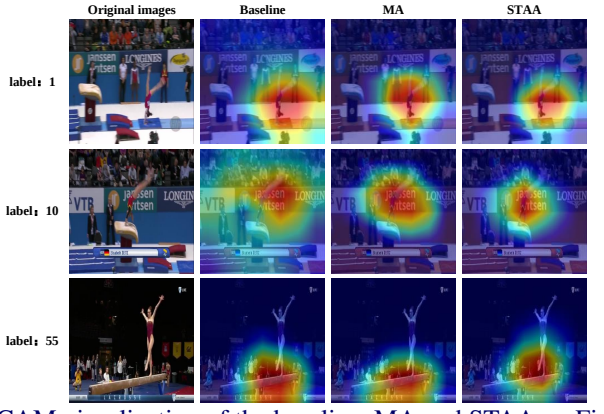

Figure 3: GradCAM visualization of the baseline, MA and STAA on FineGym dataset.

4.4.2 Effectiveness of Contrastive Learning Sampling Strategies

In Table 5, we compare the performance of different negative sampling strategies on FineGym dataset. The results indicate that Class-aware sampling achieves the highest accuracy, while random sampling and hard sample sampling yield similar results. However, the combination of random and class-aware sampling significantly outperforms others, improving accuracy by 2.1%.

Table 5: Comparison of different negative sampling strategies. Notations: S_h: hard sample, S_r: random sample, S_c: Class-aware sample.

Sampling			FineGym
S_h	S_r	S_c	
✓			93.6
	✓		93.6
		✓	93.8
✓		✓	93.7
	✓	✓	**94.0**

4.5 Visualization Analysis

In Figure 3, we present a visualization on FineGym gymnastics videos using Grad-CAM [42]. We compare the baseline and each module method separately. We observe that when utilizing only MA, the model predominantly attends to key human joint regions in the RGB image, which are highly correlated with action representation. Upon integrating TA, STAA leverages an enlarged receptive field to further enhance discriminative capacity, thereby enabling the capture of more fine-grained action features.

5 Conclusion

In this paper, we propose a Multimodal Feature Collaboration and Fusion method. The RGB modality employs a skeleton-guided cropping combined with STAA for efficient spatial and temporal feature extraction, while the skeleton modality introduces a Class-aware contrastive learning to enhance discrimination among similar actions. We then integrate the optimized RGB and skeleton features through a score-weighted fusion mechanism, achieving fine-grained feature aggregation. In future work, we plan to introduce text prompts at both the action and sub-action levels to further guide multimodal learning.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant 62406171, 62225601, U23B2052, 62306031), and in part by the Beijing Natural Science Foundation Project No. L242025.

References

[1] Ahmed Abdelkawy, Asem Ali, and Aly Farag. Epam-net: An efficient pose-driven attention-guided multimodal network for video action recognition. *Neurocomputing*, 633:129781, 2025.

[2] Dasom Ahn, Sangwon Kim, and Byoung Chul Ko. Star++: Rethinking spatio-temporal cross attention transformer for video action recognition. *Applied Intelligence*, 53(23): 28446–28459, 2023.

[3] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *CVPR*, pages 469–478, 2018.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[5] XB Bruce, Yan Liu, and Keith CC Chan. Multimodal fusion via teacher-student network for indoor action recognition. In *AAAI*, volume 35, pages 3199–3207, 2021.

[6] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022.

[7] Dongqi Cai, Yangyuxuan Kang, Anbang Yao, and Yurong Chen. Ske2grid: skeleton-to-grid representation learning for action recognition. In *ICML*, pages 3431–3441, 2023.

[8] Dongliang Chang, Yixiao Zheng, Zhanyu Ma, Ruoyi Du, and Kongming Liang. Fine-grained visual classification via simultaneously learning of multi-regional multi-grained features. *arXiv preprint arXiv:2102.00367*, 2021.

[9] Dongliang Chang, Yujun Tong, Ruoyi Du, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. An erudite fine-grained visual classification model. In *CVPR*, pages 7268–7277, 2023.

[10] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *CVPR*, pages 13359–13368, 2021.

[11] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*, pages 183–192, 2020.

[12] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *CVPR*, pages 20186–20196, 2022.

[13] Hu Cui, Renjing Huang, Ruoyu Zhang, and Tessai Hayama. Dstsa-gcn: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. *arXiv preprint arXiv:2501.12086*, 2025.

[14] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *ECCV*, pages 72–90. Springer, 2020.

[15] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9703–9717, 2021.

[16] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200*, 2017.

[17] Yifeng Ding, Zhanyu Ma, Shaoguo Wen, Jiyang Xie, Dongliang Chang, Zhongwei Si, Ming Wu, and Haibin Ling. Ap-cnn: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Transactions on Image Processing*, 30:2826–2836, 2021.

[18] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *ECCV*, pages 153–168. Springer, 2020.

[19] Ruoyi Du, Jiyang Xie, Zhanyu Ma, Dongliang Chang, Yi-Zhe Song, and Jun Guo. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9521–9535, 2021.

[20] Ruoyi Du, Wenqing Yu, Heqing Wang, Ting-En Lin, Dongliang Chang, and Zhanyu Ma. Multi-view active fine-grained visual recognition. In *ICCV*, pages 1568–1578, 2023.

[21] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, pages 2969–2978, 2022.

[22] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, pages 4768–4777, 2017.

[23] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2019.

[24] Yueying Gao, Weiguo Lin, Junfeng Xu, Wanshan Xu, and Peibin Chen. Self-supervised adversarial training for robust face forgery detection. In *BMVC*, page 718, 2023.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[26] Xiaohu Huang, Hao Zhou, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jingdong Wang, Xinggang Wang, Wenyu Liu, and Bin Feng. Graph contrastive learning for skeleton-based action recognition. *arXiv preprint arXiv:2301.10900*, 2023.

[27] Yongle Huang, Haodong Chen, Zhenbang Xu, Zihan Jia, Haozhou Sun, and Dian Shao. Sefar: Semi-supervised fine-grained action recognition with temporal perturbation and learning stabilization. *arXiv preprint arXiv:2501.01245*, 2025.

[28] Sangwon Kim, Dasom Ahn, and Byoung Chul Ko. Cross-modal learning with 3d deformable attention for action recognition. In *ICCV*, pages 10265–10275, 2023.

[29] Diksha Kurchaniya and Sanjay Kumar. Eagle: Enhanced activity recognition in occluded environments with improved gain and bi-lstm. *IEEE Transactions on Computational Social Systems*, 2025.

[30] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *ICCV*, pages 10444–10453, 2023.

[31] Mei Chee Leong, Hui Li Tan, Haosong Zhang, Liyuan Li, Feng Lin, and Joo Hwee Lim. Joint learning on the hierarchy representation for fine-grained human action recognition. In *ICIP*, pages 1059–1063. IEEE, 2021.

[32] Jianan Li, Xuemei Xie, Qingzhe Pan, Yuhan Cao, Zhifu Zhao, and Guangming Shi. Sgm-net: Skeleton-guided multimodal network for action recognition. *Pattern Recognition*, 104:107356, 2020.

[33] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019.

[34] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *CVPR*, pages 2363–2372, 2023.

[35] Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. *arXiv preprint arXiv:2411.18941*, 2024.

[36] Yalong Liu, Chengwu Liang, Songqi Jiang, and Peiwang Zhu. Video human action recognition based on motion-tempo learning and feedback attention. *Applied Sciences*, 15(8):4186, 2025.

[37] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020.

[38] Zhanyu Ma, Dongliang Chang, Jiyang Xie, Yifeng Ding, Shaoguo Wen, Xiaoxu Li, Zhongwei Si, and Jun Guo. Fine-grained vehicle classification with channel max pooling modified cnns. *IEEE Transactions on Vehicular Technology*, 68(4):3224–3233, 2019.

[39] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, pages 122–132, 2020.

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[41] Dominick Reilly and Srijan Das. Just add ! pose induced video transformers for understanding activities of daily living. In *CVPR*, pages 18340–18350, 2024.

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[43] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016.

[44] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020.

[45] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing 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*, pages 12026–12035, 2019.

[46] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[47] Sijie Song, Jiaying Liu, Yanghao Li, and Zongming Guo. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29:3957–3969, 2020.

[48] Yaohui Sun, Weiyao Xu, Xiaoyi Yu, Ju Gao, and Ting Xia. Integrating vision transformer-based bilinear pooling and attention network fusion of rgb and skeleton features for human action recognition. *International Journal of Computational Intelligence Systems*, 16(1):116, 2023.

[49] Tuyen Tran, Thao Minh Le, Duy Hung Tran, and Truyen Tran. Unified compositional query machine with multimodal consistency for video-based human activity recognition. In *BMVC*. BMVA, 2024. URL https://papers.bmvc2024.org/0426.pdf.

[50] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016.

[51] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018.

[52] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 399–417, 2018.

[53] Xuanhan Wang, Yan Dai, Lianli Gao, and Jingkuan Song. Skeleton-based action recognition via adaptive cross-form learning. In *ACM MM*, pages 1670–1678, 2022.

[54] Zhe Wang, Hao Chen, Xinyu Li, Chunhui Liu, Yuanjun Xiong, Joseph Tighe, and Charless Fowlkes. Sscap: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. In *WACV*, pages 1819–1828, 2022.

[55] Jijie Wu, Dongliang Chang, Aneeshan Sain, Xiaoxu Li, Zhanyu Ma, Jie Cao, Jun Guo, and Yi-Zhe Song. Bi-directional feature reconstruction network for fine-grained few-shot image classification. In *AAAI*, volume 37, pages 2821–2829, 2023.

[56] Shuai Xu, Dongliang Chang, Jiyang Xie, and Zhanyu Ma. Grad-cam guided channel-spatial attention module for fine-grained visual classification. In *2021 IEEE 31st international workshop on machine learning for signal Processing*, pages 1–6. IEEE, 2021.

[57] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[58] Siyuan Yang, Jun Liu, Shijian Lu, Er Meng Hwa, Yongjian Hu, and Alex C Kot. Self-supervised 3d action representation learning with skeleton cloud colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):509–524, 2023.

[59] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *CVPR*, pages 14905–14915, 2024.

[60] Haosong Zhang, Mei Chee Leong, Liyuan Li, and Weisi Lin. Pgvt: Pose-guided video transformer for fine-grained action recognition. In *WACV*, pages 6645–6656, 2024.

[61] Siqing Zhang, Ruoyi Du, Dongliang Chang, Zhanyu Ma, and Jun Guo. Knowledge transfer based fine-grained visual classification. *arXiv preprint arXiv:2012.11389*, 2020.

[62] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *CVPR*, pages 10608–10617, 2023.

[63] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *CVPR*, pages 2049–2058, 2024.

[64] Jiagang Zhu, Wei Zou, Liang Xu, Yiming Hu, Zheng Zhu, Manyu Chang, Junjie Huang, Guan Huang, and Dalong Du. Action machine: Rethinking action recognition in trimmed videos. arxiv 2018. *arXiv preprint arXiv:1812.05770*.

[65] Xiaoguang Zhu, Ye Zhu, Haoyu Wang, Honglin Wen, Yan Yan, and Peilin Liu. Skeleton sequence and rgb frame based multi-modality feature fusion network for action recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3):1–24, 2022.