Supplementary Material for MMAD: Multi-label Micro-Action Detection in Videos

Kun Li[1], Pengyu Liu[1], Dan Guo[1,2]*, Fei Wang[1], Zhiliang Wu[3], Hehe Fan[3], Meng Wang*
[1]School of Computer Science and Information Engineering, Hefei University of Technology
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3] ReLER, CCAI, Zhejiang University
kunli.hfut@gmail.com, guodan@hfut.edu.cn, eric.mengwang@gmail.com

Overview

This supplementary material provides more details of the MMA-52 dataset, methodology implementation, ablation studies, and visualization results. These topics are organized as follows.

- §1: Dataset Details.
- §2: Implementation Details.
- §3: More Experiments.
- §4: More Visualization Results.

1. Dataset Details

The **M**ulti-label **M**icro-**A**ction-52 (**MMA-52**) dataset is designed for multi-label micro-action detection. Each action instance contains both `Body-level` and `Action-level` categories. `Body-level` category denotes the body part of micro-action occurring, including ["A: Body", "B: Head", "C: Upper limb", "D: Lower limb", "E: Body-hand", "F: Head-hand", and "G: Leg-hand".] `Action-level` category denotes the exact name of micro-actions. Taking the body-level category "A: Body" as an example, there are 5 action-level categories: "A1: Shaking body", "A2: Turning around", "A3: Sitting straightly", "A4: Shrugging" and "A5: Rising up". Body-level labels and action-level labels are naturally hierarchical structures. In summary, there are **7** `Body-level` categories and **52** `Action-level` categories. The detailed descriptions and label ID of each micro-action are the same as the MA-52 dataset [2].

2. Implementation details

As shown in Figure A1, we give the detailed arches of the proposed **D**ual-path **S**patial-**T**emporal **A**dapter (**DSTA**). For the spatial-path, we use the depth-wise spatial convolution with the kernel size of 1×1 to capture the variation of micro-actions between consecutive frames. For the temporal-path, inspired by the bottleneck design in [3, 6], we design the depth-wise temporal convolution with the kernel size of $\{1, 3, 1\}$. After that, we use two separate parameters, β, and γ, to dynamically fuse the information from the spatial path and temporal path, respectively. The loss optimization of the proposed method is the same as the baseline model [5].

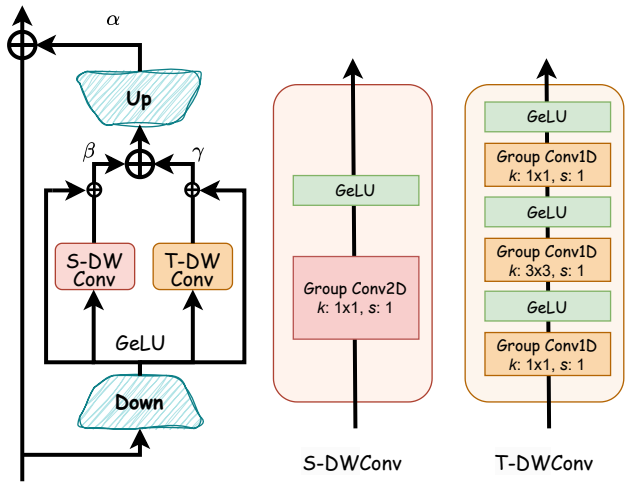


Figure A1. **Overview of the proposed Dual-path Spatial-Temporal Adapter.** "`S-DWConv`" denotes the spatial-path while "`T-DWConv`" denotes the temporal-path

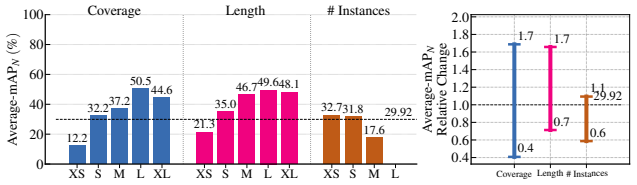
2.1. Evaluation Metric

Note that the evaluated methods can only predict the `action-level` categories. To make a fair comparison, we follow the common practice [1, 2, 4] directly mapping the predicted `action-level` category to `body-level` category. Then, we use the mAP [5] as the evaluation metric to evaluate the performance of the model on the `action-level` category and `body-level` category. To evaluate the performance at the action level and the body

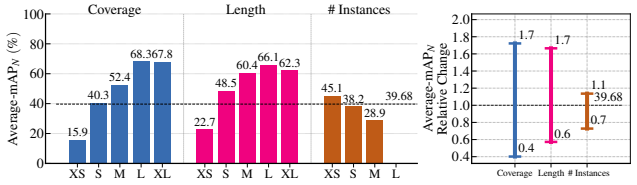
*Corresponding author

Table A1. **Ablation study of the position of the adapter.**

Layers			Action-level Detection-mAP			
1-4	5-8	9-12	@0.2	@0.5	@0.7	Avg.
✓			16.71	10.38	5.31	10.38
	✓		22.07	14.11	7.51	13.83
		✓	21.36	13.54	7.47	13.52
✓	✓	✓	**28.05**	**20.40**	**9.03**	**18.16**



(a) `Action-level` Results



(b) `Body-level` Results

Figure A2. **Sensitivity Analysis.** The left is the normalized mAP at tIoU=0.5, while the right is the relative normalized mAP change at tIoU=0.5.

level simultaneously, we report the average value (`AVG`) derived from the mean of the body-level and action-level scores.

3. More Experiments

3.1. More Ablation Studies

Ablation studies on the position of the adapter. Here, we evaluate the effect of the position of the adapter. As shown in Table A1, we divided the 12 layers of VideoMAE-S into three stages, namely early (1-4), middle (5-8), and latter (9-12). The results show that the adapter on the middle and latter layers produces better results than the early layers. The adapter on all layers achieves the best result of 18.16 in average mAP.

3.2. Sensitive Analysis

As shown in Figure A2, we perform the sensitivity analysis of the proposed model. The Average-mAP results are categorized into different buckets by three characteristics, *i.e.*, "Coverage", "Length", and "#Instances". "Coverage" denotes the ratio of instances within the video, "Length" represents the duration (seconds) of instances, and "#Instances" is the number of instances. Following the False Negative analysis settings in the main paper, the range of these characteristics are as follows, *i.e.*, "Coverage" refers

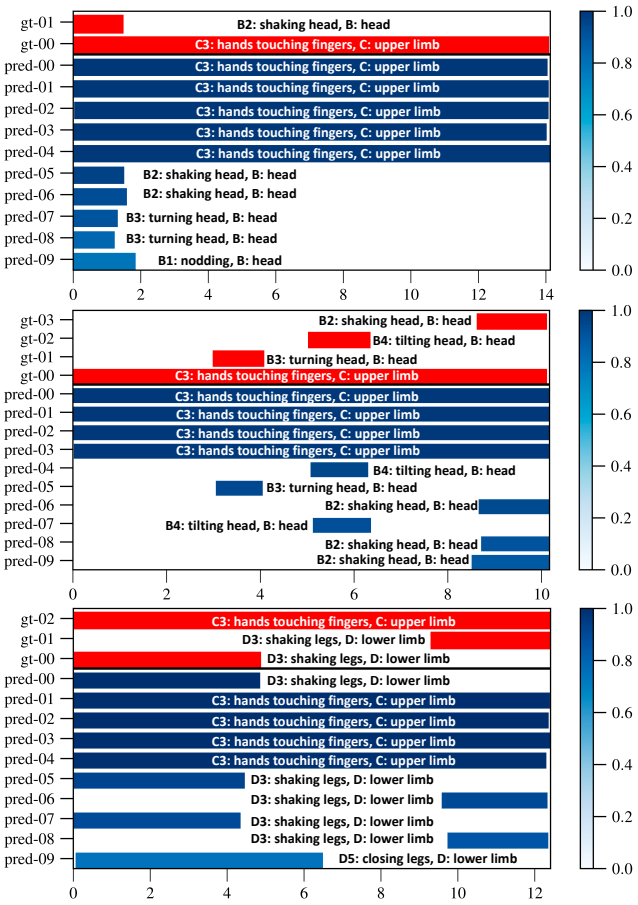


Figure A3. **Top-10 proposals of the proposed method.** In the vertical axis, gt-i represents the i-th ground truth, while pred-i denotes the i-th micro-action proposal. The color bar denotes the tIoU between the proposal and ground truth actions.

to [0.0, 0.2, 0.4, 0.6, 0.8, 1.0], "Length" refers to [0, 3, 5, 8, 9, INF], and "#Instances" refers to [-1, 2, 5, 7, INF]. In each subplot, the left column represents the mAP$_N$ and average mAP$_N$ for each specific bucket on the MMAD dataset when tIoU=0.5, while the right graph summarizes the left graph by displaying the sensitivity range, calculated as the difference between the maximum and minimum mAP$_N$ values.

From the left column, we can find that: 1) In the characteristics of coverage and length, the average mAP$_N$ gradually increases. The lowest performance is the XS bucket and the highest is the L bucket. These results suggest that instances with more coverage or longer length are generally better recognized; 2) In the characteristics of "#Instances", the average mAP$_N$ gradually decreases with the increase of action instances. The worst result is in the L bucket. These results indicate that dense instances will lead to lower performance. 3) The average mAP$_N$ at the body-level is better than the action-level, which suggests that the detection of action-level labels is easier. From the right column, we have

drawn that "Coverage" and "Length" are the most sensitive factors, with a relative change of 1.7, while "#Instances" is the least sensitive factor, with a relative change of 1.1. These results indicate the number of instances has a small influence on performance.

In summary, future work should focus on improving the accuracy of recognizing the instances with shorter coverage and length, as these are the most sensitive factors. In addition, the Average-mAP scores at the body-level are generally higher than the action-level, implying that the body-level information can be used to enhance the detection accuracy at the action-level.

4. More Visualization Results

As shown in Figure A3, we give more prediction results on the proposed MMA-52 dataset. In each subfigure, the red bars denote the ground truth instances, and the blue bars below them are the top 10 predicted proposals. The right color bar is the tIoU value between the proposal and the ground truth. The darker the color, the higher the value of tIoU. These results demonstrate that the proposed method achieves high precision in detecting micro-actions, even those with high co-occurrence and short durations.

References

[1] Jihao Gu, Kun Li, Fei Wang, Yanyan Wei, Zhiliang Wu, Hehe Fan, and Meng Wang. Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. 1

[2] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252, 2024. 1

[3] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[4] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang. Prototypical calibrating ambiguous samples for micro-action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4815–4823, 2025. 1

[5] Shuming Liu, Chen-Lin Zhang, Chen Zhao, and Bernard Ghanem. End-to-end temporal action detection with 1b parameters across 1000 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18591–18601, 2024. 1

[6] Enrique Sanchez, Mani Kumar Tellamekala, Michel Valstar, and Georgios Tzimiropoulos. Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9074–9084, 2021. 1