MMAD: Multi-label Micro-Action Detection in Videos

Kun Li[1], Pengyu Liu[1], Dan Guo[1,2]*, Fei Wang[1,2], 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

Abstract

*Human body actions are an important form of non-verbal communication in social interactions. This paper specifically focuses on a subset of body actions known as micro-actions, which are subtle, low-intensity body movements with promising applications in human emotion analysis. In real-world scenarios, human micro-actions often temporally co-occur, with multiple micro-actions overlapping in time, such as concurrent head and hand movements. However, current research primarily focuses on recognizing individual micro-actions while overlooking their co-occurring nature. To address this gap, we propose a new task named **M**ulti-label **M**icro-**A**ction **D**etection (**MMAD**), which involves identifying all micro-actions in a given short video, determining their start and end times, and categorizing them. Accomplishing this requires a model capable of accurately capturing both long-term and short-term action relationships to detect multiple overlapping micro-actions. To facilitate the MMAD task, we introduce a new dataset named **M**ulti-label **M**icro-**A**ction-**52** (**MMA-52**) and propose a baseline method equipped with a dual-path spatial-temporal adapter to address the challenges of subtle visual change in **MMAD**. We hope that MMA-52 can stimulate research on micro-action analysis in videos and prompt the development of spatio-temporal modeling in human-centric video understanding. The proposed MMA-52 dataset is available at: https://github.com/VUT-HFUT/Micro-Action.*

1. Introduction

Human body actions, as an important form of non-verbal communication, effectively convey emotional information in social interactions [2]. Previous research primarily focused on interpreting classical expressive emotions through facial expressions [11, 20, 46, 64], speech [22, 38, 75], or expressive body gestures [3, 9, 25, 26, 42]. In contrast, our study shifts the focus to a specific subset of body actions known as Micro-Actions (MAs) [9, 15, 17, 18, 28, 39, 42]. MAs are imperceptible non-verbal behaviors characterized by low-intensity movement with potential applications in emotion analysis.

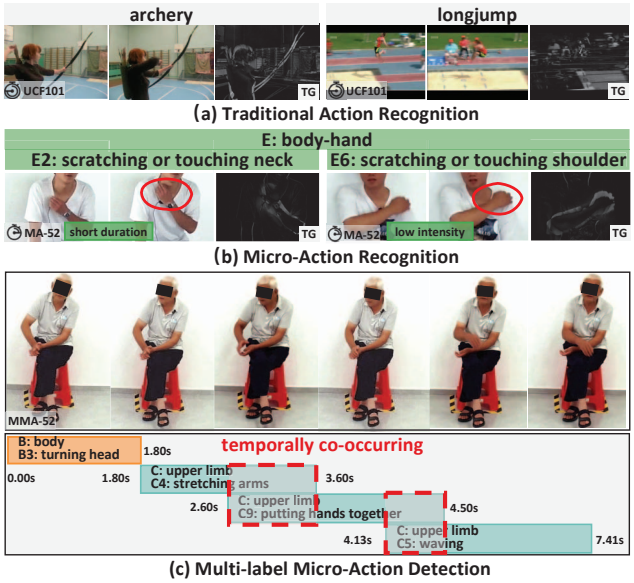


Figure 1. (a) Traditional Action Recognition [7, 24, 53] primarily focuses on actions involving large and observable movements. (b) Micro-Action Recognition [17, 28, 42] targets fine-grained actions at `body-level` and `action-level`, characterized by short duration, low intensity, and minor difference. Temporal Gradient (TG) [67] is used to visualize the subtle changes. (c) **M**ulti-label **M**icro-**A**ction **D**etection (**MMAD**) aims to detect all micro-actions within a short video, accounting for temporal co-occurrence.

Compared to traditional actions [7, 12, 13, 24, 29, 32, 53, 62], MAs have distinct characteristics as follows: **(1) Short duration.** As shown in Fig. 1, MAs typically last only a few seconds, exhibiting subtle visual changes between consecutive frames. For instance, "touching neck"

*Corresponding author

only exhibits minor changes in the neck region between a few frames. In contrast, conventional actions typically last around 5–10 seconds and involve larger and more dynamic motions in hundreds of frames. For example, the movements of "archery" or "jump" involve large motions. **(2) Low intensity.** MAs are characterized by minor spatial distinctions. As shown in Fig. 1 (b), the difference between "touching neck" and "touching shoulder" varies only in the specific contact regions involved. In contrast, conventional actions usually with identifiable motion patterns, such as those in "longjump" in Fig. 1 (a), where the overall movement is more visually distinct. **(3) Fine-grained categories.** MAs demand classification at both the body-part and action levels, involving isolated movements of individual body parts (*e.g.*, "head", "upper limb", and "lower limb") as well as coordinated motions combining parts (*e.g.*, "head-hand," and "body-hand"). In contrast, conventional action recognition typically focuses on larger-scale, whole-body motions.

Remarkable progress has been made in the micro-action recognition [15, 17, 28, 31, 59] task with the advancement of Vision Transformers [27, 30, 60, 61, 65, 66]. However, in the real world, micro-actions are **naturally temporal co-occurring**, which poses challenges for traditional action recognition methods. As shown in Fig. 1 (c), different micro-actions may occur simultaneously, such as "stretching arms" frequently happening simultaneously with "putting hands together". Therefore, driven by this intuition, we propose a new task named Multi-label Micro-Action Detection (**MMAD**) that recognizes all the micro-actions in the video sequence, achieving a fine-grained understanding of micro-actions. MMAD involves identifying all micro-actions within the video and determining their corresponding start and end times, as well as their categories. **Firstly**, MMAD requires a model capable of capturing both long-term and short-term action relationships to locate multi-scale micro-actions. **Secondly**, the model also needs to explore the complex inter-relationships between different micro-actions to ensure comprehensive detection of all possible micro-actions. **Finally**, due to the inherent nature of short duration and subtle movements in micro-actions, there is also a greater challenge in recognizing the correct categories.

To facilitate this research, we collect the first large-scale **M**ulti-label **M**icro-**A**ction-**52** (**MMA-52**) dataset, which consists of 6,528 (∼6.5k) videos with 19,782 (∼20k) action instances from 203 subjects. We first evaluate 10 baselines for traditional action detection on the MMA-52 dataset, including multi-label action detection methods and conventional temporal action detection methods. Next, we propose a baseline that incorporates a dual-path spatial-temporal adapter to capture the subtle visual changes between frames and model the associations between different actions. Specifically, the designed dual-path spatial-temporal adapter consists of two parts. In spatial, we use a depth-wise 2D convolution to model the subtle changes between adjacent frames. In temporal, we apply 1D temporal depth-wise convolution to aggregate temporal information. Finally, we use two learnable parameters to fuse temporal and spatial features separately. Extensive experiments and error analyses are conducted on the proposed benchmark dataset to validate the effectiveness of the proposed method.

Overall, the main contributions of this paper are summarized as follows:

- We introduce the task of multi-label micro-action detection (MMAD) and collect the multi-label micro-action-52 (MMA-52) dataset to facilitate the research of micro-action analysis.
- We propose an initial solution with a dual-path spatio-temporal adapter to model subtle discriminative motions. Experimental results on the benchmark dataset validate the effectiveness of the proposed method.
- We evaluate 10 baselines from the conventional temporal action detection on MMAD and in-depth studies, which reveal the inherent challenges in multi-label micro-action detection.

2. Related Work

2.1. Micro-Actions Recognition

Micro-Actions (MAs) [8, 9, 17, 28, 42] are an important form of non-verbal communication, which are usually related to humans' emotional status [2]. To facilitate the study of these subtle movements, several datasets have been constructed. To advance the study of these subtle movements, several datasets have been developed. iMiGUE [42] and SMG [9] focused on spontaneous micro-gestures in the upper limbs of athletes, revealing deep emotional states conveyed through these micro-gestures. In contrast, MPI-IGI [3] primarily examined subtle upper-body behaviors in group interactions. To better analyze and understand the whole-body movement, Guo *et al*. [17] proposed a large-scale micro-action dataset named Micro-Action (MA-52), which consists of 52 action-level MAs within 7 body-level in whole-body. They also evaluated conventional action recognition methods, including 2D CNN based [14, 35, 63], 3D CNN based [6, 57], GCN-based [47, 70], and Transformer-based [48]. More recently, Li *et al*. [28] proposed a prototypical calibrating ambiguous network, designed to mitigate the influence of the inherent ambiguity of micro-actions in micro-action recognition.

2.2. Temporal Action Detection

Temporal action detection (TAD) [10, 36, 37, 40, 54] aims to localize and classify actions in untrimmed video sequences. There have been many benchmarks focused on

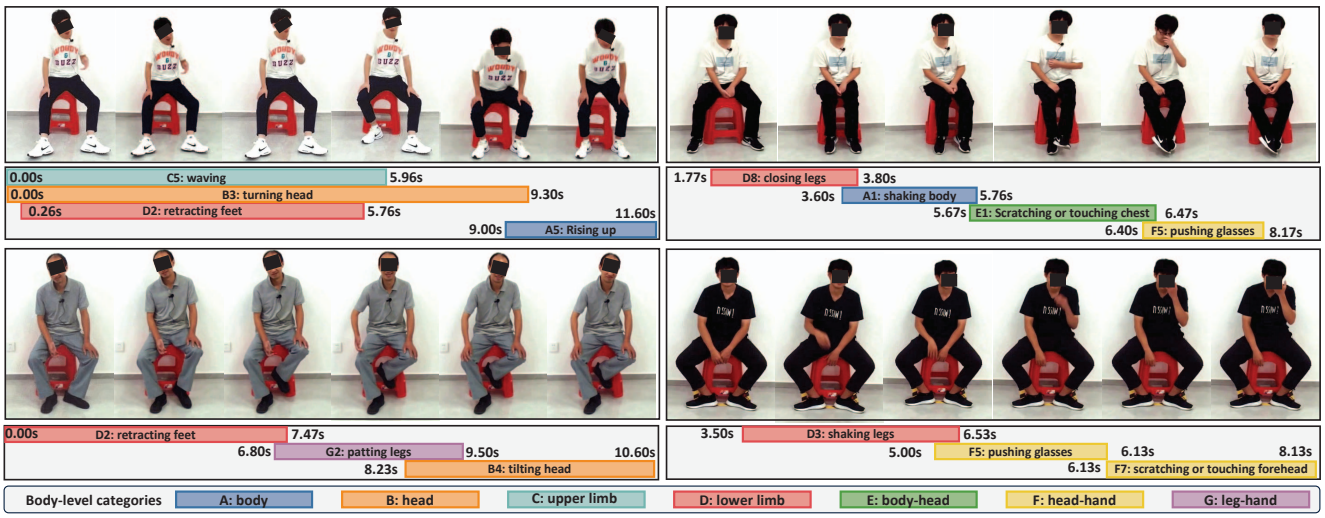


Figure 2. **Video samples from the MMA-52 dataset.** For each sample, there are different micro-action that occurs at the same time, increasing the challenge of identifying accurate micro-actions.

different domains, such as sports (THUMOS14 [23] and FineGym [51]), kitchen activities (MPII Cooking [49] and EPIC-Kitchens [50]), and daily events (ActivityNet [5], HACS Segment [74], and FineAction [45]). Driven by these datasets, TAD has witnessed significant progress, leading to the emergence of advanced methods. These approaches can be broadly categorized into feature-based methods [10, 33, 36, 37, 68, 72], which rely on pre-extracted features to detect actions, and end-to-end learning-based methods [34, 40, 43, 54, 58], which directly process raw video inputs for action localization and classification. *However*, micro-action detection is still in its infancy due to the lack of large-scale datasets. Micro-action detection remains in its early stages due to the absence of large-scale datasets. The most relevant datasets for our research are iMiGUE [42] and SMG [9], which focus on upper limb micro-gesture detection. Unfortunately, these datasets are not publicly accessible due to privacy issues. Compared to these datasets, MMA-52 surpasses existing benchmarks in terms of category diversity, number of subjects, and action instances. MMA-52 also features hierarchical labels, enabling more precise identification of multi-level MAs. We hope our MMA-52 can facilitate the research community to build robust algorithms for micro-action detection.

3. The MMA-52 dataset

3.1. Dataset Construction

Data Collection. The proposed Multi-label Micro-Action-52 (MMA-52) dataset is built upon the MA-52-Pro dataset collected by [17]. However, MA-52-Pro is not directly applicable to micro-action detection tasks due to the following reasons: ***1) Lack of fine-grained annotations***: MA-52-Pro does not provide detailed start/end timestamps for individual action instances. ***2) Variation in video lengths***: The videos in MA-52-Pro vary significantly in length, each video contains 1 to 15 MAs with durations ranging from 5s to beyond 100s. Such substantial imbalance in the action instances makes it unsuitable for action detection tasks. To address these challenges, **first**, we segment the videos into smaller clips, each ranging from 5 to 15 seconds, to ensure more consistency in video sequence and action instances. **Next**, we annotate each micro-action instance with its corresponding categories and precise start/end timestamps.

Data Annotation. Considering the inherent hierarchical nature of micro-actions [17, 28], each micro-action instance are annotated with *body-level* and *action-level* labels. In practice, annotating the multi-label micro-actions was a challenging and time-consuming task, as different types of MAs can occur simultaneously at any given moment, as illustrated in Fig. 1 (c). To ensure the accuracy of these annotations, we implemented three key measures to maintain the quality of the dataset. ***1) Annotator training***: Given the diversity of micro-action categories and the subtle differences between actions, we began by training the annotators. They first gained a thorough understanding of the definitions of the micro-action categories. Following this, we randomly selected 50 samples from each category in the micro-action recognition dataset (MA-52 [17]) and asked the annotators to perform trial annotations. Feedback based on reference annotations was provided to correct any errors arising from misunderstandings. This step ensured the annotators were well-prepared before starting the actual annotation task. ***2) Individual annotations***: Each video segment is labeled independently by three trained annotators. For each action instance, both `body-level` and `action-level` cate-

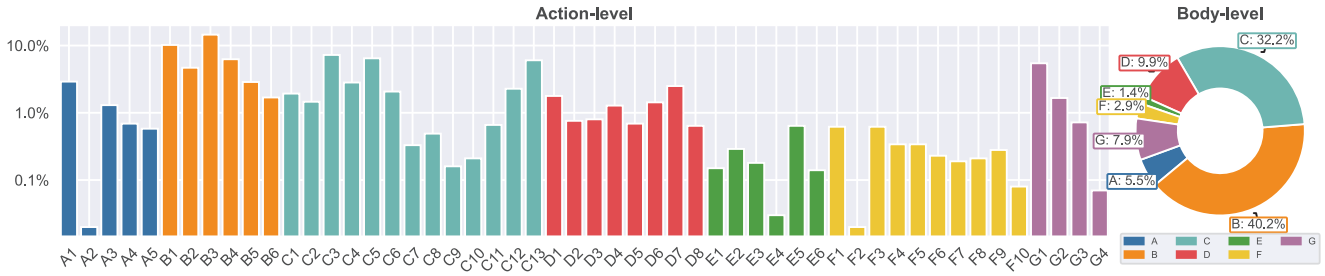


Figure 3. **The ratio of each action instance category in the MMA-52 dataset.** "A1, A2, . . . , G4" denote the `action-level` categories while "A, B, . . . , G" denote the `body-level` categories. The distribution of micro-action instances across the `action-level` and `body-level` follows a long-tail pattern. The definition of each category is the same as that of the MA-52 dataset [17].

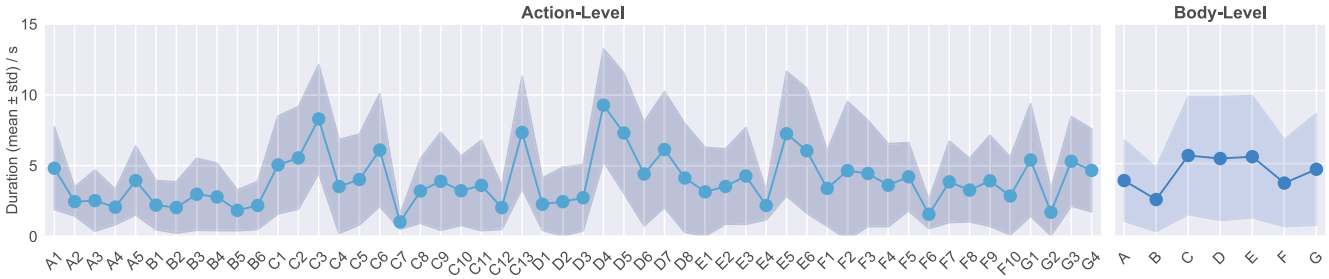


Figure 4. **The duration of each action instance in the MMA-52 dataset.** At the `action-level`, the duration of different actions varies significantly, with certain actions (*e.g.*, C3, D4, E6) exhibiting notably longer durations, while others remain considerably shorter. In contrast, the `body-level` shows a relatively lower standard deviation, indicating that body-level labels tend to be more stable.

gories are assigned, along with the corresponding start and end times. ***3) Cross-check***: After the individual annotation is completed, a cross-check is performed. If the temporal intersection-over-union (tIoU) of the same action instances annotated by all three annotators is greater than 0.9, the annotation is considered reliable. For any inconsistencies, the three annotators will discuss and follow the majority decision as the final result. This process ensures the accuracy and consistency of the annotations.

Data Partition. Since the same micro-action may exhibit individual differences, we utilize a subject-independent data partitioning strategy. As shown in Table 1, the training, validation, and test sets consist of different individuals, ensuring that no individual appears in more than one set. This strategy helps to better evaluate the model's performance across diverse subjects and enhances its generalization ability on unseen data.

3.2. Dataset Statistics and Properties

Table 1 presents the data statistics of the MMA-52 dataset, which consists of 6,528 videos, each ranging from 5 to 15 seconds in duration. The dataset contains a total of 19,782 full-body micro-action instances across 52 distinct action categories. On average, each video includes 3.1 action instances, with each instance lasting approximately 4.07 seconds. Fig. 3 presents the proportion of each action category across the three subsets. The MMA-52 dataset includes long and short micro-actions with varying temporal durations and transitions between different action states. As shown in Fig. 2, we illustrate some video samples from the MMA-52 dataset.

Table 1. **Data statistics for the MMA-52 dataset.** "Duration" refers to the length of all videos, "Avg. Video" denotes the average length of videos, and "Avg. Insta." represents the average length of action instances. "#Subj." denotes the number of subjects.

Split	Videos	Instances	Duration	Avg. Video	Avg. Insta.	#Subj.
Training	4,534	13,698	12.91h	10.25s	4.10s	140
Validation	1,475	4,735	4.34h	10.60s	4.10s	37
Test	519	1,349	1.42h	9.86s	3.79s	26
All	6,528	19,782	18.67h	10.30s	4.07s	203

Based on the above data statistics, the characteristics of the proposed MMA-52 dataset can be summarized as follows. ***1) Long-tail category distribution.*** As shown in Fig. 3, the dataset exhibits a long-tail distribution, where some MA occur frequently while others rarely occur. For example, "B3: Turning head" accounts for nearly 13% whereas "A2: Turning around" appears not up to 1%. This imbalance presents a challenge for models to enhance their robustness and ability to generalize in rare MAs. **2)** ***Variability in micro-action hierarchies.*** As shown in Fig. 4, micro-action duration exhibits significant variability, par-

Table 2. **Comparison with related datasets in micro-action analysis.** "#C" denotes the number of categories. "#Subj." denotes the number of subjects. "H-L" denotes the hierarchy action label.

Dataset	#C	#Subj.	H-L	Video	Instance	Duration	Task	Public
PAVIS F-T [4]	2	64	✗	64	N/A	N/A	Recognition	✗
MPIIGI [3]	15	78	✗	7,905	7,905	2.13s	Recognition	✓
iMiGUE [42]	32	72	✗	359	18,499	2.55s	Recognition	✗
SMG [9]	16	40	✗	414	3,712	2.14s	Recognition	✗
MA-52 [17]	52	205	✓	22,422	22,422	1.97s	Recognition	✓
iMiGUE [42]	32	72	✗	359	18,499	2.55s	Detection	✗
SMG [9]	16	40	✗	414	3,712	2.14s	Detection	✗
MMA-52 (Ours)	**52**	**203**	✓	**6,528**	**19,782**	4.07s	Detection	✓

ticularly at the action level, where some actions (*e.g.*, C3, D4, E6) last considerably longer than others. The high standard deviation suggests substantial fluctuations in duration across individuals and scenarios, posing challenges for model learning and prediction. In contrast, the body-level shows a relatively lower standard deviation, indicating that body-level labels are more stable. This highlights the need for models to capture hierarchical relationships, as individual actions are influenced by different body parts, making it insufficient to rely solely on global features. ***3) Subject-independent evaluation.*** Given the subtle nature of micro-actions, and micro-action patterns will be different across individuals, the MMA-52 dataset adopts a subject-independent setting. This means that the dataset includes action instances from a diverse range of subjects, ensuring that there is no overlap of subjects between the training, validation, and test sets. The goal is to ensure that the model learns to identify and generalize micro-actions across varying body types and movements.

3.3. Comparison with Existing Datasets

We first review the related datasets [3, 4, 9, 42] in micro-action recognition. PAVIS F-T [4] and MPIIGI [3] focus on body behaviors in group social interactions. The former concentrates solely on face-touching versus non-face-touching, while the latter analyzes a broader range of behavior categories (*e.g.*, scratch and shrug). In contrast, iMiGUE [42] and SMG [9] target upper limb micro-gestures. However, these datasets are relatively limited in terms of category diversity and subjects. To address this limitation, MA-52 [17] introduces a large-scale micro-action recognition dataset with 52 categories. Then, we compare with micro-action detection datasets. iMiGUE and SMG can be applied to action detection tasks, but they are limited to action categories. In contrast, our MMA-52 benefits from the hierarchy label (*i.e.*, `body-level` and `action-level`), large-scale action instances involving diverse subjects, enabling the design of more comprehensive and scalable detection models.

4. Methodology

4.1. Problem Formulation

The Multi-label Micro-Action Detection (MMAD) can be formulated as a set prediction problem of micro-action instances. Let the video be V containing T frames. The annotation of video V is composed by a set of micro-action instances $\Psi = \{\varphi = (t_n^s, t_n^e, c_n)\}_{N=1}^{N_g}$, where N_g is the number of micro-action instances, t_n^s and t_n^e are the starting and ending timestamp of the n-th micro-action instance, c_n is its micro-action category. The model $\mathcal{F}$ is required to predict a set of micro-action proposals $\hat{\Psi} = \{\hat{\varphi} = (\hat{t_n^s}, \hat{t_n^e}, \hat{c_n})\}_{N=1}^{N_p}$, where $\hat{t_n^s}$ and $\hat{t_n^e}$ are the predicted starting and ending timestamp of the n-th micro-action instance, $\hat{c_n}$ is its predicted micro-action category. N_p is the number of predicted micro-action instances.

4.2. Preliminary

Before introducing the baseline, we first briefly review the related techniques used in this paper.

Vanilla Adapter. As illustrated in Fig. 5 (a), the vanilla adapter [21] comprises a down-projection and an up-projection fully connected (FC) layer, with a non-linear activation function σ (such as GeLU [19]) applied between the two projections. Subsequently, a residual connection is applied to the output of the projection layer. This process can be formulated as follows:

$$\boldsymbol{X}' = \texttt{Adapter}(\boldsymbol{X}) = \boldsymbol{W}_{up}^{\top} \cdot \sigma(\boldsymbol{W}_{down}^{\top} \cdot \boldsymbol{X}) + \boldsymbol{X}, \quad (1)$$

where $\boldsymbol{W}_{down} \in \mathbb{R}^{d \times \frac{d}{r}}$ and $\boldsymbol{W}_{up} \in \mathbb{R}^{\frac{d}{r} \times d}$ denote the parameter of down- and up-projection, respectively. r is a downsampling ratio greater than one.

AdaTAD [40]. As shown on the left in Fig. 5, there is the pipeline of AdaTAD (Adapter fine-tuning for temporal action detection) [41]. An adapter is inserted into the backbone layers (*e.g.*, VideoMAE [56]) to fine-tune the model for action detection. Since the standard adapter is limited to adapting channel information, AdaTAD proposes a Temporal-Information Adapter (TIA) designed to aggregate informative local context from neighboring frames, enhancing temporal action detection. As shown in Fig. 5 (b), the temporal-informative adapter inserts a temporal depth-wise convolution (denoted as `T-DWConv`) with the kernel size of $k \times 1 \times 1$ and group size of r for depth-wise convolution to model the local contexts between adjacent frames. The TIA module is inserted into the layers between different backbones to realize transfer learning for efficient parameter tuning. Overall, the TIA module can be formulated

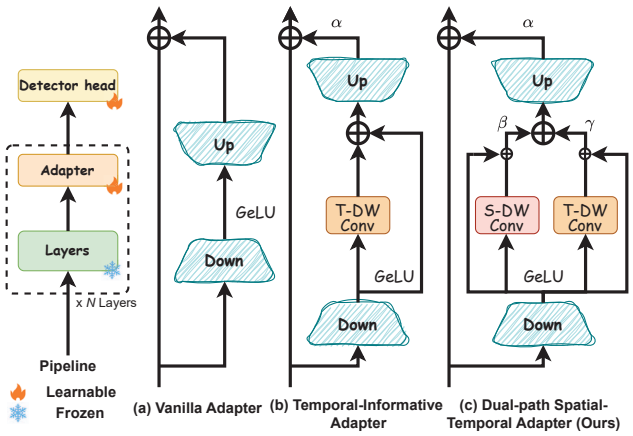


Figure 5. **Action detection pipeline and architecture of different adapters.** (a) The vanilla adapter [21] adopts the bottleneck structure. (b) The baseline model [40] designed a temporal-informative adapter to aggregate temporal information. (c) The proposed dual-path spatial-temporal adapter aggregates both spatial and temporal information.

as follows:

$$\boldsymbol{X}' = \texttt{TIA}(\boldsymbol{X}) \Leftrightarrow \begin{cases} \bar{\boldsymbol{X}} = \sigma(\boldsymbol{W}_{down}^{\top} \cdot \boldsymbol{X}), \\ \hat{\boldsymbol{X}} = \underbrace{\texttt{T-DWConv}(\bar{\boldsymbol{X}})}_{\text{temporal contexts}} + \bar{\boldsymbol{X}}, \\ \boldsymbol{X}' = \alpha \cdot \boldsymbol{W}_{up}^{\top} \cdot \hat{\boldsymbol{X}} + \boldsymbol{X}, \end{cases} \tag{2}$$

where α is a learnable parameter.

4.3. Dual-path Spatial-Temporal Adapter

As stated in the introduction, micro-action involves the subtle visual difference between adjacent frames and short duration. The baseline AdaTAD only utilizes the temporal depth-wise convolution layers to aggregate temporal information, but we argue that this will limit the model from capturing the discriminative spatial features of micro-actions. Therefore, we propose a simple Dual-path Spatial-Temporal Adapter (DSTA) to model spatial changes and temporal correlations separately. Then, the learned features from the spatial path and temporal path are fused based on the learnable weights. The above process can be formulated as follows:

$$\boldsymbol{X}' = \texttt{DSTA}(\boldsymbol{X}) \Leftrightarrow \begin{cases} \bar{\boldsymbol{X}} = \sigma(\boldsymbol{W}_{down}^{\top} \cdot \boldsymbol{X}), \\ \hat{\boldsymbol{X}}_s = \underbrace{\texttt{S-DWConv}(\bar{\boldsymbol{X}})}_{\text{spatial contexts}} + \bar{\boldsymbol{X}}, \\ \hat{\boldsymbol{X}}_t = \underbrace{\texttt{T-DWConv}(\bar{\boldsymbol{X}})}_{\text{temporal contexts}} + \bar{\boldsymbol{X}}, \\ \boldsymbol{X}' = \alpha \cdot \boldsymbol{W}_{up}^{\top} \cdot [\beta \cdot \hat{\boldsymbol{X}}_s; \gamma \cdot \hat{\boldsymbol{X}}_t] + \boldsymbol{X}, \end{cases} \tag{3}$$

where $[;]$ denotes the concatenation operation, β and γ are two learnable parameters to balance the weights of spatial and temporal pathways, respectively. `S-DWConv` symbols the spatial depth-wise convolution with the kernel size of 1×1.

5. Experiments

5.1. Experiments Setup

Evaluation Metrics. We use the mean Average Precision (mAP) [40, 54, 72] to evaluate the performance of multi-label micro-action detection. mAP measures the completeness of predicted action instances. We report the average mAP results for tIoU thresholds ranging from 0.1 to 0.9 in increments of 0.1 and the average mAP at each specific threshold. Considering the hierarchy of micro-actions, we report Detection-mAP at both `body-level` and `action-level`. Additionally, we also report the average value of mAP (`AVG`) of these two levels.

Implementation Details. We conduct experiments with the open-source toolbox OpenTAD [41]. The model employs mixed-precision training and activation checkpointing to reduce memory usage. Following [40], we use ActionFormer [72] as the detector head, retaining the original hyperparameter settings. The backbone's learning rate remains fixed, while the adapter's learning rate varies between 1e-4 and 4e-4. By default, frame resolution is set to 160^2. For the body-level prediction, we follow the common practice in micro-action analysis [16, 17] by converting the action-level results to the body-level.

5.2. Main Results

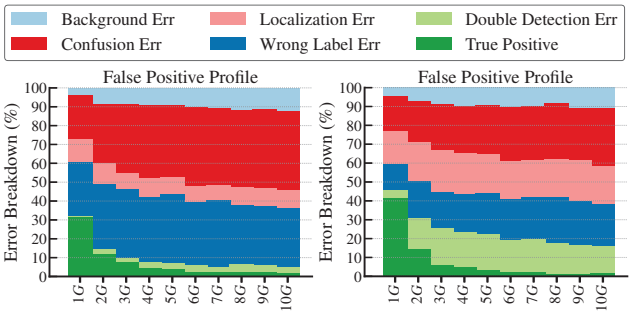
Since Multi-label Micro-Action Detection (MMAD) is a new task, we evaluate 10 baselines in conventional temporal action detection. The results are reported in Table 3. **(1)** For the multi-label temporal action detection methods (MS-TCT [10] and PointTAD [54]), these methods achieve the lowest performance. We attribute the significant performance drop to the gap between conventional actions and micro-actions. **(2)** For the feature-based methods, TemporalMaxer [55] get the best average mAP of 20.34, while the TriDet [52] only achieves the average mAP of 16.04. **(3)** For the feature-based methods, the baseline method AdaTAD [41] exhibits better performance with the backbone model enlarged from the VideoMAE small version to the base version. The best result in average mAP is 21.80, which is better than all feature-based methods. Compared to the baseline, we can see that the proposed method exhibits consistent improvement on different backbones. Specifically, on the VideoMAE-S and VideoMAE-B, there are 1.63%, and 2.26% improvements on average mAP, respectively. Although the proposed method achieves the highest

Table 3. **The experimental results on the MMA-52 dataset.** The results are measured by Detection-mAP (%) at different tIoU thresholds. The first block represents the multi-label action detection methods. The second block contains the feature-based methods for TAD, while the third block is end-to-end training methods. The best results are marked in **bold**.

Method	Backbone	Action-level				Body-level				AVG
		@0.2	@0.5	@0.7	Avg	@0.2	@0.5	@0.7	Avg	
MS-TCT [10] [CVPR2021]	I3D	5.72	3.91	2.16	3.51	12.28	8.72	4.50	7.76	5.64
PointTAD [54] [NeurIPS2021]	I3D	9.46	3.79	1.02	4.51	24.35	11.06	3.35	12.12	8.32
ActionFormer [72] [ECCV2022]	VideoMAEv2-g	23.81	16.87	8.50	15.30	40.51	24.44	12.20	23.99	19.65
TemporalMaxer [55] [arXiv2023]	VideoMAEv2-g	25.61	17.09	7.04	15.17	43.94	26.48	11.98	25.51	20.34
TriDet [52] [CVPR2023]	VideoMAEv2-g	22.41	12.06	4.59	12.45	35.60	19.99	7.84	19.62	16.04
DyFadet [71] [ECCV2024]	VideoMAEv2-g	22.17	15.19	7.96	14.17	42.52	23.34	12.55	24.93	19.55
VideoMamba [69] [ECCV2024]	VideoMAEv2-g	25.34	17.55	6.90	15.21	43.43	25.45	11.17	24.08	20.01
TadTR [44] [TIP2022]	SlowFast-R50	16.33	9.95	6.15	8.29	32.24	18.09	8.45	18.53	13.41
Re^{2}TAL [73] [CVPR2023]	Swin-Tiny	15.36	6.67	3.69	8.10	33.54	12.78	4.96	15.98	12.04
Re^{2}TAL [73] [CVPR2023]	SlowFast-101	16.15	7.10	2.93	8.10	32.39	12.15	4.57	15.38	11.74
AdaTAD [40] [CVPR2024]	VideoMAE-S	24.94	16.78	10.93	16.25	45.51	27.90	7.52	27.35	21.80
AdaTAD [40] [CVPR2024]	VideoMAE-B	28.73	19.23	8.78	17.44	49.05	28.86	7.84	28.71	23.08
DSTA (Ours)	VideoMAE-S	28.05	20.40	9.03	18.16	47.14	30.02	8.37	28.70	23.43
DSTA (Ours)	VideoMAE-B	**31.25**	**20.87**	**11.51**	**20.30**	**48.16**	**32.40**	**9.42**	**30.37**	**25.34**

Table 4. **Performance comparison under different adapters.**

Setting	Param.	mAP	gains
Snippet Feature	-	10.60	-
+ Full fine-tuning	20.89M	15.40	+4.80
+ Standard Adapter [21]	0.85M	15.87	+5.27
+ TIA (AdaTAD [40])	0.96M	16.25	+5.65
+ **DTSA (Ours)**	1.80M	**18.16**	**+7.56**



(a) `Action-level` Results (b) `Body-level` Results

Figure 6. **False Positive Profiling** on the MMA-52 dataset. The errors at the action-level mainly involve confusion and wrong label assignments while the body-level errors are evenly distributed.

average mAP of 25.34, it is still far away from conventional action detection [40]. These results indicate that there is still a gap in accurately identifying the micro-actions.

5.3. Ablation Studies

The ablation of the adapters. To validate the effectiveness of the proposed dual-path spatio-temporal adapter, we conduct experiments as follows: "Full fine-tuning" denotes fine-tuning the backbone, Standard Adapter [21], and "TIA" from the baseline model AdaTAD. The results are reported in Table 4. The baseline model only achieves the mAP of 16.25%, due to the neglect of crucial spatial information in micro-actions. In contrast, the proposed DSTA achieves the best results of 18.16 in terms of mAP, and there is 1.91% improvement.

5.4. Error Analysis

Following the convention practice [40, 52, 72, 73] in action detection, we use the tool [1] to analyze the results.

False Positive Profiling. As illustrated in Fig. 6, we conduct false positive profiling at tIoU=0.5. The x-axis is top-G predictions at tIoU=0.5, where G refers to the number of ground-truth instances. In the `action-level`, false positive errors are mainly concentrated in background, localization, and label errors, with background errors being the most prominent. In contrast, at the `body-level`, the proportion of correct detections (True Positive) increases significantly, and the error distribution is more balanced, showing a more stable performance. In summary, false positive errors at the `action-level` indicate uncertainty in the model's understanding of action-level MAs. The `body-level` shows higher accuracy, meaning the model can more accurately capture body movements.

False Negative Profiling. As shown in Fig. 7, we also conduct the False Negative Profiling under different characteristics. Specifically, "Coverage" denotes the ratio of instances within the video, "Length" represents the duration (seconds) of instances, and "#Instances" is the number of instances. Taking into account the statistics of the MMA-52

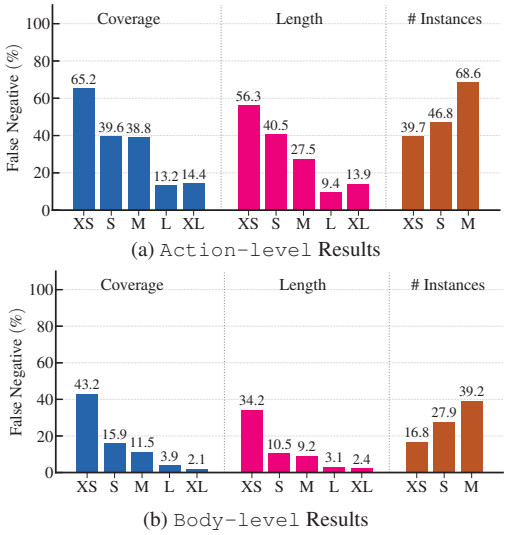


Figure 7. **False Negative Profiling** on the MMA-52 dataset. The false negatives are primarily found in instances with short coverage and duration.

dataset, the range of these characteristics are as follows, *i.e.*, "Coverage" refers 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]. These characteristic buckets are labeled as [XS, S, M, L, XL] on the axis. Based on the dimensions of Coverage and Length, we can see that False Negative rates are very high in the XS, S, and M buckets at the action-level, while only in the XS bucket at the body-level. These results suggest that detecting shorter micro-action instances remains a challenge, while longer instances are handled more effectively. In the dimension of "#Instances", False Negative rates are primarily observed in the M bucket, indicating that the key challenge is in videos with dense instances. Overall, the future direction for MMAD should focus on improving the detection of instances with low coverage and short length.

5.5. Visualization of Prediction

Additionally, we present the qualitative visualization of the detection results in Fig. 8. For the first sample, the action of "D3: shaking legs" only shows minor visual changes between frames and lasts almost the entire video. Meanwhile, "B2: shaking head" sometimes co-occurs. From the top 10 predicted proposals, we can see that the proposed method can identify the action boundaries and categories accurately. The second example involves the actions of "C4: stretching arms", "C5: waving" and "B3: turning head" across the "B: head" and "C: upper limb", our method can also detect these co-occurring actions accurately.

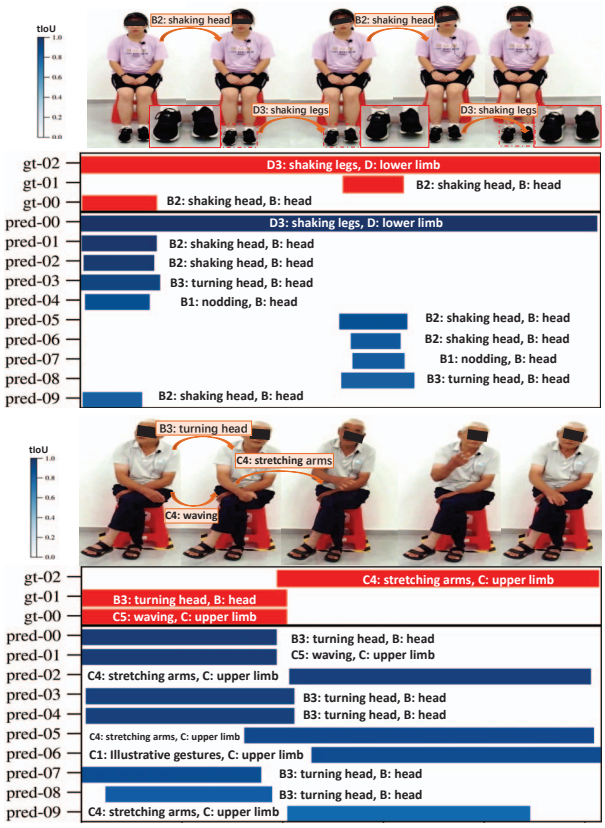


Figure 8. **Qualitative visualization of the prediction.** In the vertical axis, gt-i represents the i-th ground truth, while pred-i denotes the i-th micro-action proposal. The color bar represents the tIoU value between the proposal and the ground truth.

6. Conclusions

In this paper, we introduced the ***M**ulti-label **M**icro-**A**ction **D**etection* (***MMAD***) to tackle the challenge of identifying co-occurring micro-actions in real scenarios. To facilitate this, we developed the ***M**ulti-label **M**icro-**A**ction-**52*** (***MMA-52***) dataset, tailored for in-depth analysis and exploration of complex human micro-actions. We evaluated 10 baseline models for conventional action detection on the MMA-52 dataset. Besides, we proposed an initial solution with a dual-path spatio-temporal adapter to model the spatial variations and temporal correlations separately. The error analysis suggests that there is still a big challenge in detecting micro-actions with low coverage or short length. We hope these efforts could encourage the research community to pay more attention to the task of multi-label micro-action recognition and facilitate new advances in human body behavior analysis.

Acknowledgments

This work is supported by National Key R&D Program of China (NO.2024YFB3311602), Natural Science Foundation of China (62272144), the Anhui Provincial Natural Science Foundation (2408085J040), Anhui Provincial Key Research and Development Project (202304a05020068), the Major Project of Anhui Provincial Science and Technology Breakthrough Program (202423k09020001), the Fundamental Research Funds for the Central Universities (JZ2024HGTG0309, JZ2024AHST0337), and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *Proceedings of the European Conference on Computer Vision*, pages 256–272, 2018. 7

[2] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229, 2012. 1, 2

[3] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and Francois Bremond. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the ACM International Conference on Multimedia*, pages 70–79, 2022. 1, 2, 5

[4] Cigdem Beyan, Matteo Bustreo, Muhammad Shahid, Gian Luca Bailo, Nicolo Carissimi, and Alessio Del Bue. Analysis of face-touching behavior in large scale social interaction dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 24–32, 2020. 5

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 3

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[7] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1

[8] Guoliang Chen, Fei Wang, Kun Li, Zhiliang Wu, Hehe Fan, Yi Yang, Meng Wang, and Dan Guo. Prototype learning for micro-gesture classification. *arXiv preprint arXiv:2408.03097*, 2024. 2

[9] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023. 1, 2, 3, 5

[10] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S Ryoo, and François Brémond. Ms-tct: multi-scale temporal convtransformer for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20051, 2022. 2, 3, 6, 7

[11] Muhterem Dindar, Sanna Järvelä, Sara Ahola, Xiaohua Huang, and Guoying Zhao. Leaders and followers identified by emotional mimicry during collaborative learning: A facial expression recognition study on emotional valence. *IEEE Transactions on Affective Computing*, 13(3):1390–1400, 2020. 1

[12] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):1–18, 2018. 1

[13] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14204–14213, 2021. 1

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 2

[15] Fan Gong, Jialiang Chen, Jiajun Zhu, Qijian Bao, Fei Gao, Renshu Gu, and Gang Xu. Micro-action recognition via hierarchical fusion and inference. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 11327–11332, 2024. 1, 2

[16] 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. 6

[17] 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, 2, 3, 4, 5, 6

[18] Dan Guo, Xiaobai Li, Kun Li, Haoyu Chen, Jingjing Hu, Guoying Zhao, Yi Yang, and Meng Wang. Mac 2024: Micro-action analysis grand challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 11304–11305, 2024. 1

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[20] Saurabh Hinduja, Shaun Canavan, and Lijun Yin. Recognizing perceived emotions from facial expressions. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 236–240, 2020. 1

[21] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799, 2019. 5, 6, 7

[22] Shreyah Iyer, Cornelius Glackin, Nigel Cannings, Vito Veneziano, and Yi Sun. A comparison between convolutional and transformer architectures for speech emotion

recognition. In *2022 International Joint Conference on Neural Networks*, pages 1–8, 2022. 1

[23] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 3

[24] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563, 2011. 1

[25] Kun Li, Dan Guo, Guoliang Chen, Feiyang Liu, and Meng Wang. Data augmentation for human behavior analysis in multi-person conversations. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9516–9520, 2023. 1

[26] Kun Li, Dan Guo, Guoliang Chen, Xinge Peng, and Meng Wang. Joint skeletal and semantic embedding loss for micro-gesture classification. *arXiv preprint arXiv:2307.10624*, 2023. 1

[27] Kun Li, Dan Guo, and Meng Wang. Vigt: proposal-free video grounding with a learnable token in the transformer. *Science China Information Sciences*, 66(10):202102, 2023. 2

[28] 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, 2, 3

[29] Kun Li, Xinge Peng, Dan Guo, Xun Yang, and Meng Wang. Repetitive action counting with hybrid temporal relation modeling. *IEEE Transactions on Multimedia*, 2025. 1

[30] Qiankun Li, Xiaolong Huang, Zhifan Wan, Lanqing Hu, Shuzhe Wu, Jie Zhang, Shiguang Shan, and Zengfu Wang. Data-efficient masked video modeling for self-supervised action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2723–2733, 2023. 2

[31] Qiankun Li, Xiaolong Huang, Huabao Chen, Feng He, Qiupu Chen, and Zengfu Wang. Advancing micro-action recognition with multi-auxiliary heads and hybrid loss optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 11313–11319, 2024. 2

[32] Tianjiao Li, Lin Geng Foo, Qiuhong Ke, Hossein Rahmani, Anran Wang, Jinghua Wang, and Jun Liu. Dynamic spatio-temporal specialization learning for fine-grained action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 386–403, 2022. 1

[33] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11499–11506, 2020. 3

[34] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2021. 3

[35] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2

[36] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 2, 3

[37] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 2, 3

[38] Na Liu, Yuan Zong, Baofeng Zhang, Li Liu, Jie Chen, Guoying Zhao, and Junchao Zhu. Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5144–5148, 2018. 1

[39] Pengyu Liu, Fei Wang, Kun Li, Guoliang Chen, Yanyan Wei, Shengeng Tang, Zhiliang Wu, and Dan Guo. Micro-gesture online recognition using learnable query points. *arXiv preprint arXiv:2407.04490*, 2024. 1

[40] 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. 2, 3, 5, 6, 7

[41] Shuming Liu, Chen Zhao, Fatimah Zohra, Mattia Soldan, Alejandro Pardo, Mengmeng Xu, Lama Alssum, Merey Ramazanova, Juan León Alcázar, Anthony Cioppa, Silvio Giancola, Carlos Hinojosa, and Bernard Ghanem. Opentad: A unified framework and comprehensive study of temporal action detection. *arXiv preprint arXiv:2502.20361*, 2025. 5, 6

[42] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10631–10642, 2021. 1, 2, 3, 5

[43] Xiaolong Liu, Song Bai, and Xiang Bai. An empirical study of end-to-end temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20010–20019, 2022. 3

[44] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022. 7

[45] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing*, 31: 6937–6950, 2022. 3

[46] Yang Liu, Xingming Zhang, Janne Kauttonen, and Guoying Zhao. Uncertain facial expression recognition via multi-task

assisted correction. *IEEE Transactions on Multimedia*, 26: 2531–2543, 2024. 1

[47] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020. 2

[48] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 2

[49] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201, 2012. 3

[50] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119:346–373, 2016. 3

[51] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2020. 3

[52] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 6, 7

[53] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[54] Jing Tan, Xiaotong Zhao, Xintian Shi, Bin Kang, and Limin Wang. Pointtad: Multi-label temporal action detection with learnable query points. In *Advances in Neural Information Processing Systems*, pages 15268–15280, 2022. 2, 3, 6, 7

[55] Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023. 6, 7

[56] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022. 5

[57] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 2

[58] Chenhao Wang, Hongxiang Cai, Yuxin Zou, and Yichao Xiong. Rgb stream is enough for temporal action detection. *arXiv preprint arXiv:2107.04362*, 2021. 3

[59] Chen Wang, Xun Mei, and Feng Zhang. Instance-aware fine-grained micro-action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 11320–11326, 2024. 2

[60] Fei Wang, Dan Guo, Kun Li, and Meng Wang. Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5345–5353, 2024. 2

[61] Fei Wang, Dan Guo, Kun Li, Zhun Zhong, and Meng Wang. Frequency decoupling for motion magnification via multi-level isomorphic architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18984–18994, 2024. 2

[62] Fei Wang, Kun Li, Yiqi Nie, Zhangling Duan, Peng Zou, Zhiliang Wu, Yuwei Wang, and Yanyan Wei. Exploiting ensemble learning for cross-view isolated sign language recognition. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2453–2457, 2025. 1

[63] 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. 2

[64] Yi Wu, Shangfei Wang, and Yanan Chang. Patch-aware representation learning for facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6143–6151, 2023. 1

[65] Zhiliang Wu, Changchang Sun, Hanyu Xuan, Gaowen Liu, and Yan Yan. Waveformer: wavelet transformer for noise-robust video inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6180–6188, 2024. 2

[66] Zhiliang Wu, Kerui Chen, Kun Li, Hehe Fan, and Yi Yang. Bvinet: Unlocking blind video inpainting with zero annotations. *arXiv preprint arXiv:2502.01181*, 2025. 2

[67] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[68] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020. 3

[69] Shihao Xu, Jing Fang, Xiping Hu, Edith Ngai, Wei Wang, Yi Guo, and Victor CM Leung. Emotion recognition from gait analyses: Current research and future directions. *IEEE Transactions on Computational Social Systems*, 11(1):363–377, 2022. 7

[70] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[71] Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, and Fan Li. Dyfadet: Dynamic feature aggregation for temporal action detection. In *European Conference on Computer Vision*, pages 305–322, 2024. 7

[72] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510, 2022. 3, 6, 7

[73] Chen Zhao, Shuming Liu, Karttikeya Mangalam, and Bernard Ghanem. Re2tal: Rewiring pretrained video backbones for reversible temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10637–10647, 2023. 7

[74] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 3

[75] Jiaqi Zhao, Fei Wang, Kun Li, Yanyan Wei, Shengeng Tang, Shu Zhao, and Xiao Sun. Temporal-frequency state space duality: An efficient paradigm for speech emotion recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025. 1