Advancing Micro-Action Recognition with Multi-Auxiliary Heads and Hybrid Loss Optimization

Qiankun Li
HFIPS, Chinese Academy of Sciences
University of Science and Technology of China
Hefei, China
qklee@mail.ustc.edu.cn

Xiaolong Huang
Mila - Quebec AI Institute
Concordia University
Montreal, Canada
hirox827@gmail.com

Huabao Chen
HFIPS, Chinese Academy of Sciences
University of Science and Technology of China
Hefei, China
hbchen98@gmail.com

Feng He
University of Science and Technology of China
Hefei, China
hefengcs@gmail.com

Qiupu Chen
HFIPS, Chinese Academy of Sciences
University of Science and Technology of China
Hefei, China
qpuchen@mail.ustc.edu.cn

Zengfu Wang*
HFIPS, Chinese Academy of Sciences
University of Science and Technology of China
Hefei, China
zfwang@ustc.edu.cn

Abstract

Video action recognition has been a hot research direction in computer vision, with most existing technologies focusing on coarse-grained macro-action recognition. However, fine-grained action recognition remains challenging. Micro-actions, characterized by high fine-grained, low-intensity, and brief, are crucial for emotion recognition and psychological assessment applications. In this paper, we build on popular video action recognition frameworks as foundation models, introducing multi-auxiliary heads and hybrid loss optimization to advance micro-action recognition. Specifically, the Frame-Level pred and Coarse-Grained Body-Action auxiliary heads work collaboratively to enhance the model and Fine-Grained Micro-Action primary head for perceiving fine-grained and capturing keyframes. Incorporating F1 loss, ArcFace loss, and weighted multi-task loss improves training stability, convergence speed, and performance. Additionally, integrating the optical flow modality enriches the model's diversity, and ensemble learning across all foundational models. Finally, our method achieves a 75.37% F1-mean on the MA-52 dataset, ranking 1st in the Micro-Action Analysis Grand Challenge in conjunction with ACM MM'24. The code is available at https://github.com/qklee-lz/ACMMM2024-MAC.

CCS Concepts

• **Computing methodologies → Activity recognition and understanding**.

*Corresponding author.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3688975

Keywords

Video Micro-Action Recognition; Multi-Auxiliary Heads; Hybrid Loss Optimization; Fine-Grained Action Recognition

ACM Reference Format:
Qiankun Li, Xiaolong Huang, Huabao Chen, Feng He, Qiupu Chen, and Zengfu Wang. 2024. Advancing Micro-Action Recognition with Multi-Auxiliary Heads and Hybrid Loss Optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3664647.3688975

1 Introduction

Video action recognition has emerged as a pivotal research direction within the field of computer vision, driven by its extensive applications in intelligent surveillance [46], human-computer interaction [28], and sports analysis [31]. Despite significant advances, most existing technologies predominantly focus on coarse-grained macro-action recognition [3, 22, 23]. These methods achieve impressive performance by identifying broad, high-level activities such as running or jumping [2, 17, 33]. However, they often fall short in capturing the intricate details required for fine-grained action recognition, particularly micro-actions [10, 19].

Micro-actions are characterized by their high fine granularity, low intensity, and brevity, which present unique challenges [29]. Micro-action recognition aims to detect and distinguish ephemeral body movements, generally occurring within a temporal span of $1/25s \sim 1/3s$ [10]. These subtle movements are often critical for applications centered on emotion recognition and psychological assessment [18], where minute details can reveal significant insights into a subject's emotional and psychological state [26, 44]. The ability to accurately recognize these micro-actions holds the potential to enhance various human-centered applications [4, 25, 45].

CNN-based approaches typically employ 3D convolutions [15, 35] or temporal convolutions [30, 36, 43] to capture motion features from sequences of RGB frames implicitly. These methods excel in extracting spatiotemporal information by learning the dynamics present across consecutive frames. Transformer-based methods, such as TimeSformer [1], have revolutionized video processing by applying spatial-temporal self-attention directly to the time series, optimizing the handling of spatiotemporal data. The Video Swin Transformer [27] segments video frames into patches and processes them through a hierarchical Swin Transformer structure across different stages, achieving impressive performance in capturing fine-grained details over extended temporal windows. In addition to supervised learning methods, self-supervised learning has gained traction by designing video-specific pretraining tasks that do not require manual annotations [9, 14, 37]. Methods like Masked Video Modeling (MVM) [8, 34, 40] have shown remarkable results in learning rich video representations. Prominent examples include VideoMAE [7, 34, 38], InternVideo [41, 42], and UnMasked Teacher [20]. Moreover, the integration of multimodal information has been extensively explored to improve video action recognition. By incorporating data from optical flow [32], SlowFast networks [6], pose [22], text [41], and audio [16], these multimodal approaches provide complementary perspectives that enrich the representation and understanding of complex actions. While these advanced macro-action recognition techniques provide powerful baselines, custom developments are still required for the challenging micro-action recognition [10, 19].

In this paper, we propose a novel approach specifically designed to address the challenges of micro-action recognition. Our method builds on popular video action recognition frameworks as foundation models (VideoMAE [34], Video Swin Transformer [27], InternVideo [41], and UnMasked Teacher [20]), introducing multi-auxiliary heads combined with hybrid loss optimization to advance micro-action recognition. Specifically, we incorporate Frame-Level pred and Coarse-Grained Body-Action auxiliary heads to work collaboratively with the Fine-Grained Micro-Action primary head. The Frame-Level pred auxiliary head enhances the model's ability to capture fine-grained details and keyframes by making preds at each frame. The Coarse-Grained Body-Action auxiliary head introduces body action priors to explicitly enhance the model's understanding of coarse-grained actions, working in tandem with the Frame-Level pred head to improve fine-grained recognition implicitly. Through joint training of these auxiliary heads and the Micro-Action primary head, our model significantly improves its perception of subtle actions. To improve training stability, convergence speed, and overall performance, we integrate F1, ArcFace, and weighted multi-task loss into our optimization process. Additionally, we enrich our model's diversity by integrating the optical flow modality, and we utilize ensemble learning across all foundational models to further boost performance.

Our proposed method demonstrates its effectiveness by achieving a 75.37% F1-mean on the MA-52 dataset, securing the first position in the Micro-Action Analysis Grand Challenge [11] in conjunction with ACM MM'24. This significant improvement highlights the potential of our approach in advancing the state-of-the-art in micro-action recognition.

The contributions of this paper are summarized as follows:

- We introduce a novel framework based on several macro video action recognition foundation models for advancing micro-action recognition.
- We propose multi-auxiliary heads, including Frame-Level pred and Coarse-Grained Body-Action auxiliary heads, to enhance the perception and capture of fine-grained and keyframe details.
- A hybrid loss optimization strategy is designed, combining F1, ArcFace, and weighted multi-task loss to improve training stability, convergence speed, and performance.
- The optical flow modality is incorporated, and ensemble learning is utilized across foundational models to enrich the diversity and robustness of the model.
- Our method achieves a 75.37% F1-mean on the MA-52 dataset, ranking 1st in the Micro-Action Analysis Grand Challenge in conjunction with ACM MM'24.

2 Task Background

Micro-action recognition (MAR) is a specialized field within action recognition that focuses on identifying and classifying subtle, often overlooked movements that convey significant non-verbal information. Unlike more conspicuous actions like running or jumping, micro-actions, such as slight nods or quick leg shakes, provide deep insights into an individual's emotions and intentions, making this field particularly valuable for applications in emotion recognition and psychological assessments.

2.1 Challenges in Micro-action Recognition

One of the main challenges in MAR is the subtle and transient nature of these actions. They occur across different body parts and can be very rapid, making them difficult to capture and distinguish. Additionally, the visual similarity between different micro-actions adds another layer of complexity to their accurate identification and differentiation. Due to these characteristics, MAR tasks require highly precise and efficient algorithms capable of handling low-amplitude fluctuations in gestures and postures.

2.2 The MA-52 Dataset

The MA-52 dataset [10] is specifically designed for micro-action recognition, offering a substantial collection of 22,422 video samples. These samples are categorized into 52 fine-grained micro-action classes and 7 coarse-grained body parts, providing a rich dataset for model training and evaluation.

Consider a micro-action video defined as $\mathcal{V} = \{I_1, I_2, \ldots, I_T\}$, where T is the length of the video. The objective is to classify the micro-actions contained in the video by selecting them from a set of micro-action labels $\mathcal{Y}$. The affiliation relationship between fine-grained micro-actions $\{y_1, \ldots, y_{N_A^F}\}$ and coarse-grained body parts $\{y_1, \ldots, y_{N_A^C}\}$ is annotated, where N_A^F and N_A^C represent the number of micro-action categories and body parts, respectively. The recognized fine-grained micro-action category serves as an indicator of its corresponding coarse-grained body part.

2.3 Evaluation Metrics

The evaluation metrics for the MA-52 dataset are designed to address the unique challenges posed by micro-action recognition, including imbalanced data distribution and the need for both fine-grained and coarse-grained classification accuracy. The primary evaluation metric is the F1 score, which is used to assess the performance of models on both micro-actions and body parts.

The F1 score is calculated in both micro and macro variants to provide a comprehensive evaluation:

$$\mathrm{F1}_{\mathrm{micro}} = 2 \cdot \frac{\overline{\mathrm{Pre}} \cdot \overline{\mathrm{Recall}}}{\overline{\mathrm{Pre}} + \overline{\mathrm{Recall}}},$$

$$\overline{\mathrm{Pre}} = \frac{\sum_{i=1}^{N} \mathrm{TP}_i}{\sum_{i=1}^{N} (\mathrm{TP}_i + \mathrm{FP}_i)}, \quad \overline{\mathrm{Recall}} = \frac{\sum_{i=1}^{N} \mathrm{TP}_i}{\sum_{i=1}^{N} (\mathrm{TP}_i + \mathrm{FN}_i)},$$

$$\mathrm{F1}_{\mathrm{macro}} = 2 \cdot \frac{\mathrm{Pre} \cdot \mathrm{Recall}}{\mathrm{Pre} + \mathrm{Recall}},$$

$$\mathrm{Pre} = \frac{1}{N_C} \sum_{j=1}^{N_C} \mathrm{Pre}_j, \quad \mathrm{Recall} = \frac{1}{N_C} \sum_{j=1}^{N_C} \mathrm{Recall}_j,$$

where N is the number of samples and N_C is the number of categories. For the MA-52 dataset, these metrics are computed for both coarse- and fine-grained labels. The final evaluation metric is the mean F1 score, calculated as follows:

$$\mathrm{F1}_{\mathrm{mean}} = \frac{\mathrm{F1}_{\mathrm{macro}}^{\mathrm{coarse}} + \mathrm{F1}_{\mathrm{micro}}^{\mathrm{coarse}} + \mathrm{F1}_{\mathrm{macro}}^{\mathrm{fine}} + \mathrm{F1}_{\mathrm{micro}}^{\mathrm{fine}}}{4}.$$

This composite metric ensures a balanced assessment of the model's performance across different levels of granularity and data distributions.

3 Method

3.1 Pre-trained Foundation Models

In general video action recognition, there are many strong baselines. Selecting appropriate methods as foundation models and then customizing them for micro-action recognition tasks holds promise for achieving satisfactory results. Additionally, using multiple baseline models pre-trained on various benchmarks can enhance diversity, thereby improving the effectiveness of subsequent ensemble learning.

Supervised video action recognition. The Video Swin Transformer [27] has consistently been a powerful baseline in supervised action recognition, performing well across multiple benchmarks with rich pre-training resources. Therefore, it is suitable as one type of foundation model.

Self-supervised video representation learning. Masked Video Modeling (MVM) leverages the inherent structure and spatiotemporal information in video data to learn robust video representations without extensive labeling, thus improving model performance across various video analysis tasks. Notably, MVM enhances the model's understanding of video coherence and consistency, thereby improving action recognition. Given the advanced nature and abundant pre-training resources of these models, we adopt VideoMAE [34], InternVideo [41, 42], and UnMasked Teacher [20] as our foundation models.

Pre-training dataset. Popular large-scale public pre-trained video action recognition datasets include the standard Kinetics series (Kinetics-400, Kinetics-600, Kinetics-700, Kinetics-710), AVA-Kinetics, and Something-Something-V2 (SSV2). Experience in the era of large models shows that scaling up data volume and model size often results in stronger performance and higher accuracy in handling complex tasks [13, 21, 24]. Among these datasets, the Kinetics series is larger than the others. For our micro-action recognition task, having a video representation with more comprehensive dynamic temporal information is beneficial. However, experience in video action recognition [40] suggests that Kinetics series data tend to favor spatial cues, with static background information being sufficient for most action discrimination (e.g., play football on a green field, basketball on a court). In contrast, SSV2 and AVA-Kinetics rely more on dynamic temporal information. Therefore, each dataset has its advantages, and considering the benefits of diversity in ensemble learning, we selected the pre-trained foundation models listed in Table 1.

Table 1: Pre-trained Foundation Models.

Method	Version	Pre-train	Frames	Size
Video-Swin [27]	Tiny/Base	K400	32	224
Video-Swin [27]	Base	SSv2	32	224
VideoMAE [34]	Base	K710/SSv2	16	224
VideoMAE [34]	Large	K700	16	224
VideoMAE [34]	Huge	AVA-Kinetics	16	224
InternVideo2 [41]	1B	K700	16	224
UMT [20]	Large	K700	16	224

3.2 Multi-Auxiliary Heads

To enhance the performance of micro-action recognition, we employ multiple auxiliary heads in our model. These auxiliary heads aim to provide additional supervision and improve the learning of fine-grained and coarse-grained features. Specifically, we use a frame-level pred auxiliary head and a coarse-grained body-action auxiliary head.

Frame-level pred auxiliary head. The frame-level pred auxiliary head focuses on refining the model's ability to recognize actions at the granularity of individual frames. Let $\mathbf{X} \in \mathbb{R}^{N\times C\times T\times H\times W}$ represent the input video features, where N is the batch size, C is the number of channels, T is the number of frames, H and W are the height and width of each frame.

For each frame $t \in \{1, 2, \ldots, T\}$, the auxiliary head processes the features through a series of operations. First, the features are spatially averaged:

$$\mathbf{X}'_{nt} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_{nctij} \quad \forall\, n, c, t,$$

where $\mathbf{X}'_{nt} \in \mathbb{R}^{N\times T\times C}$ represents the temporally aggregated features. These features are then processed by a squeeze-and-excitation (SE) block $\mathrm{SE}(\cdot)$ [12] to capture channel-wise dependencies:

$$\mathbf{X}''_{nt} = \mathrm{SE}(\mathbf{X}'_{nt}).$$

Finally, the frame-level action class probabilities are predicted using a fully connected layer:

$$\mathbf{z}_{nt} = \mathrm{FC}(\mathbf{X}''_{nt}) \in \mathbb{R}^{K},$$

where K is the number of fine-grained action classes.

Coarse-grained body-action auxiliary head. The coarse-grained body-action auxiliary head focuses on recognizing broader body actions that span multiple frames, providing a body prior and higher-level understanding of the action sequence. The input features $\mathbf{X} \in \mathbb{R}^{N \times C \times T \times H \times W}$ are first pooled spatially and temporally:

$$\mathbf{G}_n = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_{nctij} \quad \forall\, n, c,$$

where $\mathbf{G}_n \in \mathbb{R}^{N \times C}$ represents the aggregated features for each video. These aggregated features are then used to predict the coarse-grained body-action class probabilities using a fully connected layer:

$$\mathbf{z}_n = \mathrm{FC}(\mathbf{G}_n) \in \mathbb{R}^{M},$$

where M is the number of coarse-grained body-action classes.

Additionally, frame-level coarse preds are also generated similarly to the fine-grained preds:

$$\mathbf{z}_{nt}^{\text{coarse}} = \mathrm{FC}_{\text{coarse}}(\mathbf{X}_{nt}'') \in \mathbb{R}^{M}.$$

Collaborating with the fine-grained micro-action primary head. The frame-level pred and coarse-grained body-action auxiliary heads collaborate with the fine-grained micro-action primary head to enhance the model's capability of perceiving fine-grained details and capturing keyframes. The primary head can be implemented using popular methods such as TimeSformerHead [1], TSN [39], or I3D [2]. The auxiliary heads are then integrated into these primary methods to enhance their performance. The implementation involves adaptive average pooling, dropout, and fully connected layers, ensuring that the model effectively captures both detailed and broad action features. This collaboration significantly boosts the overall performance of micro-action recognition tasks.

3.3 Hybrid Loss Optimization

In this section, we detail the loss functions used to optimize the model. The optimization process combines multiple loss functions to effectively handle the fine-grained and coarse-grained action classification tasks. Specifically, we use F1 loss, ArcFace loss, and a weighted multi-task loss.

F1 loss. The F1 loss is designed to address the class imbalance issue by focusing on the harmonic mean of precision and recall. Following the MA-52 evaluation metric, we introduce both macro and micro F1 losses for the fine-grained action classes. Let $\mathbf{p} = \mathrm{softmax}(\mathbf{s})$ be the predicted probability distribution, and $\mathbf{y}$ be the one-hot encoded ground truth labels. The true positives (TP), false positives (FP), and false negatives (FN) are computed as follows:

$$\mathrm{TP}_c = \sum_{i=1}^{N} \mathbf{y}_{ic} \cdot \mathbf{p}_{ic}, \quad \mathrm{FP}_c = \sum_{i=1}^{N} (1 - \mathbf{y}_{ic}) \cdot \mathbf{p}_{ic}, \quad \mathrm{FN}_c = \sum_{i=1}^{N} \mathbf{y}_{ic} \cdot (1 - \mathbf{p}_{ic}),$$

where c is the class index and N is the number of samples.

The precision and recall for each class are given by:

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \epsilon}, \quad \mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c + \epsilon},$$

where ϵ is a small constant to avoid division by zero.

The macro F1 score is the average of the F1 scores for all classes:

$$\mathrm{F1}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c + \epsilon}.$$

The micro F1 score is computed globally over all samples:

$$\mathrm{F1}_{\text{micro}} = \frac{2 \cdot \mathrm{Precision}_{\text{all}} \cdot \mathrm{Recall}_{\text{all}}}{\mathrm{Precision}_{\text{all}} + \mathrm{Recall}_{\text{all}} + \epsilon},$$

where

$$\mathrm{Precision}_{\text{all}} = \frac{\sum_{c=1}^{C} \mathrm{TP}_c}{\sum_{c=1}^{C} (\mathrm{TP}_c + \mathrm{FP}_c) + \epsilon}, \mathrm{Recall}_{\text{all}} = \frac{\sum_{c=1}^{C} \mathrm{TP}_c}{\sum_{c=1}^{C} (\mathrm{TP}_c + \mathrm{FN}_c) + \epsilon}.$$

The combined F1 loss is then:

$$\mathcal{L}_{\text{F1}} = 1 - (\mathrm{F1}_{\text{macro}} + \mathrm{F1}_{\text{micro}}).$$

ArcFace loss. ArcFace loss enhances the discriminative power of the model by optimizing the angular margin between classes. It modifies the softmax function to include an additive angular margin penalty.

Given the input features $\mathbf{x}$ and the corresponding class weight $\mathbf{W}$, the cosine similarity is calculated as:

$$\cos(\theta_j) = \frac{\mathbf{W}_j^{\top} \mathbf{x}}{\|\mathbf{W}_j\| \|\mathbf{x}\|},$$

where $\| \cdot \|$ denotes the L_2 norm.

The modified cosine similarity with the angular margin m is:

$$\cos(\theta_j + m) = \cos(\theta_j)\cos(m) - \sin(\theta_j)\sin(m).$$

The scaled and penalized cosine similarity is then:

$$s \cdot \cos(\theta_j + m),$$

where s is a scaling factor. The ArcFace loss is defined as:

$$\mathcal{L}_{\text{ArcFace}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s \cdot (\cos(\theta_{y_i} + m))}}{e^{s \cdot (\cos(\theta_{y_i} + m))} + \sum_{j \neq y_i} e^{s \cdot \cos(\theta_j)}}.$$

Weighted multi-task Loss. The final loss function combines the fine-grained and coarse-grained classification tasks with weighted contributions. Let $\mathcal{L}_{\text{fine}}$, $\mathcal{L}_{\text{coarse}}$, $\mathcal{L}_{\text{fine-frame}}$, and $\mathcal{L}_{\text{coarse-frame}}$ be the fine-grained micro-action loss, coarse-grained body-action loss, frame-level fine-grained micro-action loss, and frame-level coarse-grained body-action loss, respectively. The total loss is:

$$\begin{aligned}\mathcal{L}_{\text{total}} = {} & \lambda_{\text{fine}} \mathcal{L}_{\text{fine}} + \lambda_{\text{coarse}} \mathcal{L}_{\text{coarse}} \\ & + \lambda_{\text{fine-frame}} \mathcal{L}_{\text{fine-frame}} + \lambda_{\text{coarse-frame}} \mathcal{L}_{\text{coarse-frame}},\end{aligned}$$

where $\lambda_{\text{fine}} = 1, \lambda_{\text{coarse}} = 0.5, \lambda_{\text{fine-frame}} = 0.5$, and $\lambda_{\text{coarse-frame}} = 0.25$ are the weights for the respective losses.

Each loss is composed of both the combined F1 loss and ArcFace loss as follow:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{F1}}^{\text{fine}} + \mathcal{L}_{\text{ArcFace}}^{\text{fine}}, \quad \mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{F1}}^{\text{coarse}} + \mathcal{L}_{\text{ArcFace}}^{\text{coarse}},$$

$$\begin{aligned}\mathcal{L}_{\text{fine-frame}} = \frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(& \mathcal{L}_{\text{F1}}^{\text{frame}}(\mathbf{z}_{nt}, \mathbf{y}_{it}) \\ & + \mathcal{L}_{\text{ArcFace}}^{\text{frame}}(\mathbf{z}_{nt}, \mathbf{y}_{it}) \Big),\end{aligned}$$

$$\mathcal{L}_{\text{coarse-frame}} = \frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(\mathcal{L}_{\text{F1}}^{\text{coarse-frame}}(\mathbf{z}_{nt}^{\text{coarse}}, \mathbf{y}'_{it}) + \mathcal{L}_{\text{ArcFace}}^{\text{coarse-frame}}(\mathbf{z}_{nt}^{\text{coarse}}, \mathbf{y}'_{it}) \Big),$$

where $\mathbf{y}'_{it}$ = fine2coarse($\mathbf{y}_n$) maps fine-grained mirco-action labels to coarse-grained body-action labels.

Finally, the final weighted multi-task loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fine}} + 0.5 \cdot \mathcal{L}_{\text{coarse}} + 0.5 \cdot \mathcal{L}_{\text{fine-frame}} + 0.25 \cdot \mathcal{L}_{\text{coarse-frame}}.$$

3.4 Data Processing

Human-centered crop data augmentation strategy. To enhance the effectiveness of our model, we implement a human-centered crop data augmentation strategy. This involves using YOLO detector or OpenCV to detect humans in each frame, taking the union of the detected bounding boxes, and then expanding the bounding box slightly to ensure the completeness of the human subjects while filtering out excess background. To maintain an appropriate aspect ratio, we ensure the ratio is not less than 0.6 by adding padding to the sides of the bounding box when necessary. This prevents distortion and improves the framing of human subjects. This strategy centers and highlights the human subjects, reducing irrelevant background influence and enabling the model to focus on key actions. By maintaining a consistent aspect ratio, we provide clearer and more relevant visual information, enhancing action recognition performance.

Optical flow. We utilize the VideoMAE backbone model to extract features from both optical flow and RGB modalities, and then fuse these features through concatenation or addition to enhance model diversity and facilitate ensemble learning. Specifically, we compute dense optical flow for each video. This process involves converting frames to grayscale, calculating the optical flow using the Farneback method [5], and transforming the flow vectors into visual representations. These representations encode the magnitude and direction of motion, producing images that provide additional features for the model to recognize actions more effectively.

Pipeline. The data processing pipelines for training, validation, and testing involve several steps to ensure the input data is appropriately preprocessed. For training, the pipeline includes uniform sampling, random resizing, cropping, color jittering, random erasing, and flipping. Validation and testing pipelines also use uniform sampling and decoding, followed by resizing and center cropping to standardize the input dimensions.

3.5 Ensemble Learning

Given our multiple foundation models and diverse module designs, we employ ensemble learning to enhance performance. We use weighted averaging to combine the preds from different models. Specifically, the final pred probability $\mathbf{p}^*$ is computed as a weighted average of the individual model preds, $\mathbf{p}^* = \sum w_i \mathbf{p}_i$, where w_i are the weights assigned based on model performance, ensuring $\sum w_i = 1$.

This approach leverages the complementary strengths of each model, improving accuracy and robustness, and resulting in better generalization and reduced variance in action recognition tasks.

4 Experiment

4.1 Main Results

Main results of various foundation models. We present the main results of our experiments, comparing the performance of various foundation models and configurations on the test sets. Table 2 lists the F1 mean scores. The results indicate that models incorporating all components, such as the VideoMAE with TSN head and all components, achieve the highest performance on the test* set with an F1 mean score of 73.13%.

Ensemble learning results. As listed in Table 2, we use various foundation models with different components, exhibiting high diversity, making them well-suited for ensemble learning. We employ weighted averaging to combine the predictions from these models. The weights used are 0.7, 0.9, 0.7, 0.4, 0.4, 0.2, 0.3, 0.6, 0.4, 0.9, 0.6, 0.7, 0.8, and 0.7. The final ensemble achieves an F1 mean score of 75.37%, top-1 accuracy for body parts of 85.59%, top-1 accuracy for actions of 70.83%, F1 macro for body parts of 82.47%, F1 micro for body parts of 85.59%, F1 macro for actions of 62.58%, and F1 micro for actions of 70.83%, ranking 1st in the Micro-Action Analysis Grand Challenge in conjunction with ACM MM'24.

4.2 Ablation Study

Effect of Pre-training. We conducted an ablation study to evaluate the effect of pre-training on the performance of our models. Specifically, we used the Video-Swin Transformer model with different pre-training datasets and versions. Table 3 lists the results of the validation set.

The results demonstrate that pre-training on larger (K600 > K400) and more relevant datasets (video data > image data) significantly improves model performance. Interestingly, although the SSv2 dataset is smaller than the Kinetics series, it yields better fine-tuning results. This may be because SSv2 focuses more on dynamic temporal cues, while Kinetics emphasizes static spatial cues, the former aligns better with the requirements of micro-action recognition tasks. As dataset size increases, the performance gap between these narrows.

Effect of Data Processing. We evaluated the impact of different data processing strategies and larger scale pre-training on model performance using VideoMAE Base with TimeSformerHead pre-trained on K700. The baseline model employs random frame sampling. We investigated the effects of uniform sampling, pre-training on K710, and the human-centered crop data augmentation strategy. The results are presented in Table 4. The results indicate that all these strategies contribute to performance improvement.

Effect of F1 and ArcFace loss. Incorporating F1 loss into the baseline model improved the F1 mean score from 69.51% to 70.09%. Adding ArcFace loss also resulted in a performance increase, with the F1 mean score rising from 69.51% to 69.89%. The results indicate that both F1 and ArcFace loss contribute to better model performance in micro-action recognition tasks.

Effect of auxiliary head. We evaluated the impact of different auxiliary heads on the performance of the VideoMAE Base model with the data processing strategies, F1 loss, and ArcFace loss applied

Table 2: Performance comparison of different models and configurations. Test* indicates training with both train and val sets.

Method	Head	Version	Pre-train	Test	Test*	Components
Video-Swin	I3D	Base	SSv2	70.80	-	F1ArcFaceLoss
Video-Swin	I3D	Base	SSv2	69.48	71.54	Data Aug, F1ArcFaceLoss, Frame-level pred
Video-Swin	I3D	Large	K700	66.51	71.02	Data Aug, F1ArcFaceLoss, Frame-level pred
VideoMAE	I3D	Base	K710	71.14	72.81	Data Aug, F1ArcFaceLoss, Frame-level pred
VideoMAE	TSN	Base	K710	71.18	73.13	All Components
VideoMAE	I3D	Base	K710	71.11	-	Data Aug, F1ArcFaceLoss
VideoMAE	I3D	Base	K710	70.04	71.56	F1ArcFaceLoss, Frame-level pred, OpticalFlow (Weight Sharing)
VideoMAE	TimeSformer	Base	SSv2	68.94	-	Data Aug
VideoMAE	TimeSformer	Base	K710	-	71.94	Data Aug
VideoMAE	I3D	Base	K710	-	72.17	F1ArcFaceLoss, Frame-level pred, OpticalFlow
VideoMAE	TimeSformer	Large	K700	-	73.06	Data Aug
VideoMAE	TimeSformer	Huge	K700	-	72.68	Data Aug
UMT	TimeSformer	Large	K700	-	72.37	Data Aug
InternVideo2	TimeSformer	1B	K700	69.87	-	Data Aug

Table 3: Effect of pre-training on F1 mean performance.

Method	Version	Pre-train	F1 mean
Video-Swin	Tiny	ImageNet-1K	64.45
Video-Swin	Tiny	K400	66.19
Video-Swin	Base	K400	66.69
Video-Swin	Base	K600	67.40
Video-Swin	Base	SSv2	67.82

Table 4: Effect of data processing strategies on F1 mean performance.

Baseline	Uniform	K710	Human-centered Crop	F1 mean
✓				69.51
	✓			70.02
	✓	✓		71.20
	✓	✓	✓	72.32

(except for the baseline in the first row). Table 5 lists the results on validation and test sets.

The results indicate that incorporating frame-level and coarse-grained auxiliary heads generally improves performance. The TSN model with both auxiliary heads achieved the highest F1 mean score of 73.13%, which is the strongest single model in our method.

Effect of optical flow. We evaluated the impact of using optical flow alongside RGB modalities on the performance of the VideoMAE Base model with the frame-level pred auxiliary head. Table 6 shows the results on the validation and test sets.

The results indicate that using the optical flow modality did not lead to significant performance improvements or even a slight decrease. However, the introduction of optical flow modality can still increase diversity in subsequent ensemble learning.

Table 5: Effect of auxiliary heads on F1 mean performance (based on VideoMAE Base). Test* indicates training with both train and val sets.

MAE Head	Frame-level	Coarse-grained	Val	Test	Test*
TimeSformer			72.32	70.80	71.94
TimeSformer	✓		72.28	71.11	-
I3D	✓		72.33	71.14	72.81
I3D	✓	✓	72.58	70.98	73.07
TSN	✓	✓	72.70	71.18	**73.13**

Table 6: Effect of optical flow on F1 mean performance. Test* indicates training with both train and val sets.

Modality	Val	Test*
Optical Flow only	67.61	-
RGB only	72.33	72.81
Optical Flow + RGB	-	72.17

5 Conclusion

This paper advanced micro-action recognition by innovatively integrating multi-auxiliary heads and hybrid loss functions with foundational video action recognition frameworks. Our approach, emphasizing fine-grained recognition through Frame-Level and Coarse-Grained Body-Action heads, significantly enhanced keyframe capture and fine-grained perceptivity. The incorporation of F1 loss, ArcFace loss, and weighted multi-task loss not only stabilized training but also accelerated convergence, leading to a notable performance increase. Furthermore, the addition of optical flow as a modality and the use of ensemble learning strategies substantially enriched our model's capability. Finally, our method achieves a 75.37% F1-mean on the MA-52 dataset, ranking 1st in the Micro-Action Analysis Grand Challenge in conjunction with ACM MM'24.

References

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

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

[3] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9640–9649.

[4] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. 2023. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *arXiv preprint arXiv:2312.05447* (2023).

[5] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 363–370.

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

[7] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems* 35 (2022), 35946–35958.

[8] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10406–10417.

[9] Jie Gui, Tuo Chen, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. 2023. A Survey of Self-Supervised Learning from Multiple Perspectives: Algorithms, Theory, Applications and Future Trends. *arXiv preprint arXiv:2301.05712* (2023).

[10] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. 2024. Benchmarking Micro-action Recognition: Dataset, Methods, and Applications. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 7 (2024), 6238–6252. https://doi.org/10.1109/TCSVT.2024.3358415

[11] Dan Guo, Xiaobai Li, Kun Li, Haoyu Chen, Jingjing Hu, Guoying Zhao, Yi Yang, and Meng Wang. 2024. MAC 2024: Micro-Action Analysis Grand Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[12] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[13] Xiaolong Huang, Qiankun Li, Xueran Li, and Xuesong Gao. 2024. One Step Learning, One Step Review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 12644–12652.

[14] Yuqi Huo, Mingyu Ding, Haoyu Lu, Zhiwu Lu, Tao Xiang, Ji-Rong Wen, Ziyuan Huang, Jianwen Jiang, Shiwei Zhang, Mingqian Tang, et al. 2021. Self-supervised video representation learning with constrained spatiotemporal jigsaw. (2021).

[15] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 221–231.

[16] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5492–5501.

[17] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*. IEEE, 2556–2563.

[18] Jia Li, Jiantao Nie, Dan Guo, Richang Hong, and Meng Wang. 2022. Emotion separation and recognition from a facial expression by generating the poker face with vision transformers. *arXiv preprint arXiv:2207.11081* (2022).

[19] Kun Li, Dan Guo, Pengyu Liu, Guoliang Chen, and Meng Wang. 2024. MMAD: Multi-label Micro-Action Detection in Videos. *arXiv preprint arXiv:2407.05311* (2024).

[20] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19948–19960.

[21] Qiankun Li, Xiaolong Huang, Bo Fang, Huabao Chen, Siyuan Ding, and Xu Liu. 2023. Embracing large natural data: Enhancing medical image analysis via cross-domain fine-tuning. *IEEE Journal of Biomedical and Health Informatics* (2023).

[22] Qiankun Li, Xiaolong Huang, Yuwen Luo, Xiaoyu Hu, Xinyu Sun, and Zengfu Wang. 2023. Mitigating Context Bias in Action Recognition via Skeleton-Dominated Two-Stream Network. In *Proceedings of the 2023 Workshop on Advanced Multimedia Computing for Smart Manufacturing and Engineering*. 65–70.

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

[24] Qiankun Li, Yimou Wang, Yani Zhang, Zhaoyu Zuo, Junxin Chen, and Wei Wang. 2024. Fuzzy-ViT: A Deep Neuro-Fuzzy System for Cross-Domain Transfer Learning from Large-scale General Data to Medical Image. *IEEE Transactions on Fuzzy Systems* (2024).

[25] Qiankun Li, Xianwang Yu, Junxin Chen, Ben-Guo He, Wei Wang, Danda B Rawat, and Zhihan Lyu. 2023. PGA-Net: Polynomial Global Attention Network With Mean Curvature Loss for Lane Detection. *IEEE Transactions on Intelligent Transportation Systems* (2023).

[26] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10631–10642.

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

[28] Jianyuan Ni, Anne HH Ngu, and Yan Yan. 2022. Progressive Cross-Modal Knowledge Distillation for Human Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5903–5912.

[29] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *IEEE transactions on affective computing* 12, 2 (2018), 505–523.

[30] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.

[31] Hideo Saito, Thomas B Moeslund, and Rainer Lienhart. 2022. MMSports' 22: 5th International ACM Workshop on Multimedia Content Analysis in Sports. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7386–7388.

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

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[34] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[35] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[36] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*. 6450–6459.

[37] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. 2020. Self-supervised video representation learning by pace prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 504–521.

[38] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14549–14560.

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

[40] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. 2022. BEVT: BERT pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14733–14743.

[41] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. 2024. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377* (2024).

[42] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191* (2022).

[43] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*. 305–321.

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

[45] Sravani Yenduri, Nazil Perveen, Vishnu Chalavadi, et al. 2022. Fine-grained action recognition using dynamic kernels. *Pattern Recognition* 122 (2022), 108282.

[46] Dingwen Zhang, Chaowei Fang, Wu Liu, Xinchen Liu, Jingkuan Song, Hongyuan Zhu, Wenbing Huang, and John Smith. 2022. HCMA'22: 3rd International Workshop on Human-Centric Multimedia Analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7407–7409.