SeFAR: Semi-supervised Fine-grained Action Recognition with Temporal Perturbation and Learning Stabilization

Yongle Huang[1,2*], **Haodong Chen**[1,2*], **Zhenbang Xu**[1], **Zihan Jia**[3], **Haozhou Sun**[4], **Dian Shao**[1†]

[1]Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China
[2]School of Automation, Northwestern Polytechnical University, Xi'an, China
[3]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[4]School of Software, Northwestern Polytechnical University, Xi'an, China
{yonglehuang, chd}@mail.nwpu.edu.cn, shaodian@nwpu.edu.cn

Abstract

Human action understanding is crucial for the advancement of multimodal systems. While recent developments, driven by powerful large language models (LLMs), aim to be general enough to cover a wide range of categories, they often overlook the need for more specific capabilities. In this work, we address the more challenging task of Fine-grained Action Recognition (FAR), which focuses on detailed semantic labels within shorter temporal duration (e.g., "salto backward tucked with 1 turn"). Given the high costs of annotating fine-grained labels and the substantial data needed for fine-tuning LLMs, we propose to adopt semi-supervised learning (SSL). Our framework, SeFAR, incorporates several innovative designs to tackle these challenges. Specifically, to capture sufficient visual details, we construct Dual-level temporal elements as more effective representations, based on which we design a new strong augmentation strategy for the Teacher-Student learning paradigm through involving moderate temporal perturbation. Furthermore, to handle the high uncertainty within the teacher model's predictions for FAR, we propose the Adaptive Regulation to stabilize the learning process. Experiments show that SeFAR achieves state-of-the-art performance on two FAR datasets, FineGym and FineDiving, across various data scopes, as well as two classical coarse-grained datasets, UCF101 and HMDB51. Further analysis and ablation studies validate the effectiveness of our designs. Additionally, we show that the features extracted by SeFAR could largely promote the ability of multimodal models to understand fine-grained and domain-specific semantics.

Code — https://github.com/KyleHuang9/SeFAR

Introduction

Understanding videos has attracted increasing attention as videos contain vivid visual information and rich temporal dynamics absent in text and images. In the past year, we have seen remarkable progress in multimodal large language models (MLLMs) (Hu et al. 2024; Chen et al. 2023; Li et al. 2024, 2023b), aiming at acquiring more general and comprehensive abilities. However, as pointed out by recent studies (Zhao et al. 2024; Yuan et al. 2023), chasing generality may sacrifice some task-specific performance, which motivates us to delve into a perpendicular direction: focus on more *specific* tasks to promote the fine-grained understanding ability of models.

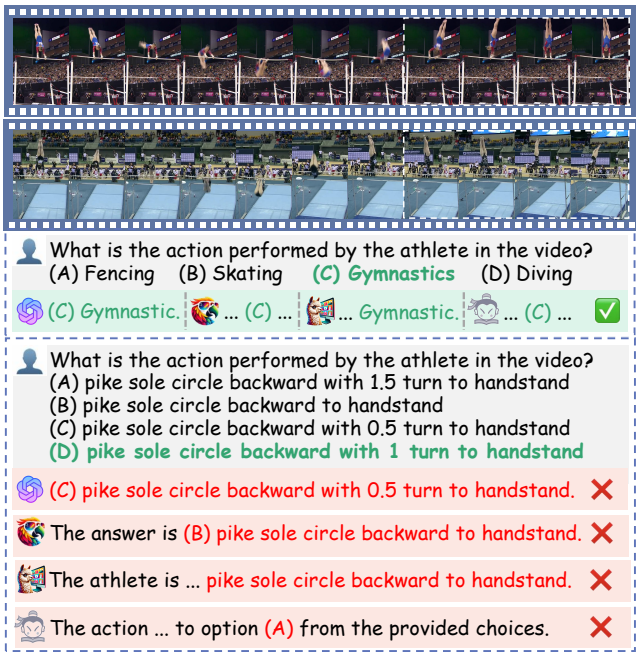

Figure 1: **Fine-grained Action Instances.** The two samples are drawn from the FineGym (Shao et al. 2020) dataset, specifically the *"pike sole circle backward with 0.5 turn to handstand"* at the top and the *"... 1 turn ..."* at the bottom. We further test popular MLLMs on the bottom instance for both coarse-grained and fine-grained: GPT-4V (OpenAI 2024), VideoChat2 (Li et al. 2024), VideoLLaVA (Lin et al. 2023), and InternLM-XComposer-2.5 (Zhang et al. 2024).

Specifically, we focus on Fine-grained Action Recognition (FAR), a challenging human-centric video understanding task. To explain, classical action recognition (Xiong et al. 2021; Xiao et al. 2022; Dave et al. 2023; Xing et al. 2023) only demands the model to provide relatively coarse-grained category such as *"gymnastics"*, while FAR aims to provide more detailed, specific, and semantically accurate

*These authors contributed equally.
†Corresponding author

descriptions as *"pike sole circle backward with 0.5 turn to handstand"*. To demonstrate the difficulty of this task, we evaluate four powerful MLLMs (OpenAI 2024; Li et al. 2024; Lin et al. 2023; Zhang et al. 2024), as shown in Fig. 1. Unfortunately, they all fail to correctly recognize the fine-grained semantics of the given action. In such a sense, FAR holds significance in further enhancing the capability of MLLM (Driess et al. 2023; Vemprala et al. 2024), especially in application scenes requiring more accurate and professional information.

However, limited research on FAR not only owes to its higher demands for method design but also the dataset construction (Shao et al. 2020; Xu et al. 2022a). For example, providing annotations such as *"5237D with 3.5 twists"* (Xu et al. 2022a) requires adequate expert knowledge, huge annotation time, and large checking efforts to ensure the quality (Shao et al. 2020). This leads to the scarcity of fine-grained labels and makes it difficult to directly re-train or fine-tune large models with huge annotated data. Keep this in mind, we further adopt the semi-supervised learning (SSL) setting, where only a small percentage of labeled data is needed (Zhu 2005). Consequently, targeting semi-supervised FAR, besides those intrinsic challenges from both sides, we have to tackle intractable *new challenges* that emerged when combined. Specifically, FAR needs enough visual details, effective information aggregation, and a comprehensive understanding of temporal dynamics (Shao et al. 2020; Xu et al. 2022a; Li et al. 2022; Tang et al. 2023). For SSL, the core is to equip the unlabeled data with stable and reasonable supervision (*e.g.*, pseudo-labels) (Sohn et al. 2020; Zhu 2005; Kurakin et al. 2020). However, when training a semi-supervised FAR model, the generated pseudo-labels may not be reliable, since FAR is rather challenging, making the whole learning process easily collapse.

In this paper, we propose a novel framework, **SeFAR**, to address the above challenges. Due to the semi-supervised setting, SeFAR is developed based on the FixMatch (Sohn et al. 2020) SSL paradigm, including the weak-to-strong consistency regularization and the Teacher-Student setup, as shown in Fig. 2. Moreover, there are also delicately designed strategies and modules incorporated in SeFAR: ❶ First, to effectively mine adequate and useful data for FAR, a *dual-level information modeling* strategy is proposed. This process combines both fine-grained temporal elements with the temporal context to effectively capture multi-granular temporal information, enhancing the ability to discriminate subtle actions in the video. ❷ Then, to construct weak-strong contrast data pairs more tailored for FAR which differs from the traditional spatial-only augmentations (Yun et al. 2019; DeVries 2017; Kurakin et al. 2020), we highlight the significance of temporal dynamics and design a new strong augmentation strategy. Specifically, we introduce *moderate temporal perturbation* into the fine-grained temporal elements achieved previously, while keeping the temporal order of context element. ❸ Moreover, in order to provide reliable pseudo-labels for unlabeled data even when the Teacher model suffers from unstable predictions, we design an *Adaptive Regulation* to stabilize the training process by calculating coefficients to adjust the losses. In addition, to directly tackle the problems outlined in Fig. 1, we adhere to the standard MLLM framework, which includes a vision encoder, a language encoder, and an alignment adapter. By incorporating our SeFAR model as an innovative video encoder, we observe that all MLLMs perform better on FAR, as shown in Tab. 5.

To summarize, our contributions are as follows:

- To the best of our knowledge, this is the first work to explore the highly challenging task of **Se**mi-supervised **F**ine-grained **A**ction **R**ecognition and an effective framework **SeFAR** is proposed for this purpose, which is based on the FixMatch paradigm but incorporates a new augmentation strategy to form the weak-to-strong data pairs;
- Moreover, SeFAR incorporates several innovative designs to address specific challenges, including the dual-level temporal elements modeling, careful involvement of moderate temporal perturbation, as well as the adaptive regulation for a steady learning process;
- SeFAR achieves state-of-the-art performance on both fine-grained (FineGym, FineDiving) and coarse-grained action recognition datasets (UCF101, HMDB51), demonstrating its effectiveness. Additional analysis shows that SeFAR could also serve as a powerful visual encoder to assist current MLLMs in domain-specific scenes.

Related Work

Fine-grained Action Recognition (FAR). FAR aims to differentiate between similar human actions at a finer semantic granularity (*e.g.*,*"switch leap with 0.5 turns" vs. "split jump with 1 turn"*), while coarse-grained actions (Zhou et al. 2018; Xu et al. 2022b; Yang et al. 2020), stop at the granularity of *"gymnastics"*. To achieve this, abundant and subtle motion details are extremely desired (Shao et al. 2020). There are several pioneer works (Li et al. 2022; Leong et al. 2022; Tang et al. 2023) to tackle the problem of FAR. However, they have predominantly focused on fully supervised or few-shot learning. Among them, LCDC (Mac et al. 2019) capture local spatio-temporal features, HAAN (Li, He, and Xu 2022) use hierarchical modeling with atomic actions and visual concepts, while M^3Net (Tang et al. 2023) implement multi-view encoding, matching, and fusion. Distinct from the above works, we propose to address a more challenging task and propose the first semi-supervised FAR framework, SeFAR, integrating with the *dual-level temporal elements modeling*, which tackles the subtle inter-class differences but also contends with limited annotations.

Data Augmentation in Semi-supervised Learning (SSL). Data augmentation plays an essential role in SSL, serving as one of the two core components of the FixMatch (Sohn et al. 2020)-based paradigm, specifically *consistency regularization* achieved through both strong and weak data augmentation. This has been previously demonstrated. For instance, (Xie et al. 2020) emphasizes that a robust model should withstand variations in input examples or hidden states. However, most existing semi-supervised video action recognition studies (Xu et al. 2022b; Xiong et al. 2021; Xiao et al. 2022; Dave et al. 2023) focus primarily on spatial augmentations achieved through image-based strategies (*e.g.*,

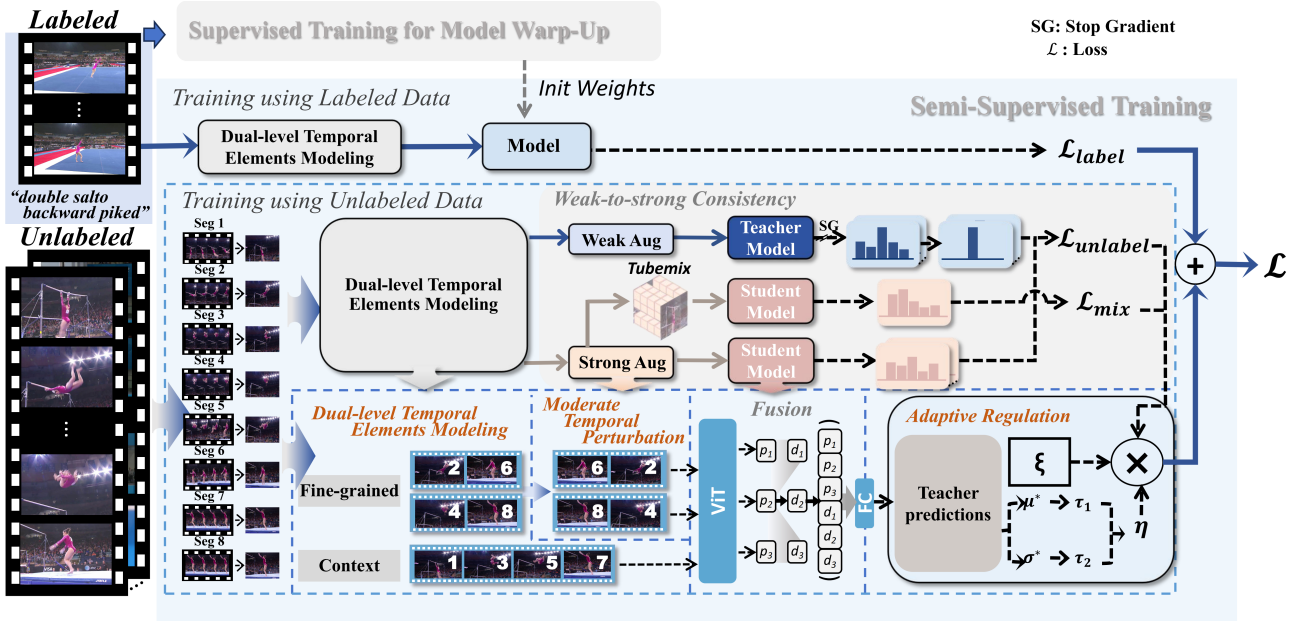

Figure 2: **Overview of SeFAR pipeline.** We target Semi-supervised FAR, assuming most input samples are unlabeled. During unsupervised learning, SeFAR adopts *dual-level temporal elements modeling* and performs augmentation in two manners ('Weak' *vs.* 'Strong'). Strongly augmented/distorted samples by *moderate temporal perturbation* are used by the student model, while the teacher model offers pseudo-labels based on weakly augmented samples. Consistency is enforced through loss minimization ($\mathcal{L}_{un}$). The unsupervised loss is further adjusted by our proposed *Adaptive Regulation*. The framework is trained with a weighted combination of supervised $\mathcal{L}_{sup}$ and unsupervised $\mathcal{L}_{un}$ losses.

Cutmix (Yun et al. 2019), Cutout (DeVries 2017), or their variants (Kurakin et al. 2020; Cubuk et al. 2020)). We argue that temporal augmentation is equally important inspired by (Xing et al. 2023), especially in FAR, as spatial augmentations can often disrupt critical information within actions. To address this, we design a new temporal augmentation strategy, *moderate temporal perturbation*. Furthermore, to maintain stability in the *pseudo-labeling* process, another core component of the FixMatch-based paradigm, we have developed the *Adaptive Regulation* during training.

Methodology

To tackle the challenging task of semi-supervised fine-grained action recognition, we propose the SeFAR framework, and the complete pipeline is shown in Fig. 2. Before delving into specific details, we first elaborate on the preliminaries about semi-supervised learning, especially the FixMatch (Sohn et al. 2020) paradigm.

Preliminaries

❑ **Teacher *vs.* Student Model.** A line of SSL frameworks adopts the Teacher-Student setting, where the *Teacher* provides pseudo-labels to supervise the *Student* model. Instead of directly sharing weights between teacher and student models (Sohn et al. 2020), we adopt an average of consecutive student models to obtain a "Mean teacher", whose effectiveness has been verified (Tarvainen and Valpola 2017). Formally, at a given time step, the weights of the *Teacher* model, θ_t, is updated as an exponential moving average of the student weights θ_s:

$$\theta_t \longleftarrow \omega\theta_s + (1-\omega)\theta_t. \tag{1}$$

As pointed out in (Xing et al. 2023), such EMA-Teacher is more suitable and stable for human action recognition.

❑ **Weak *vs.* Strong Augmentation.** One core component within FixMatch (Sohn et al. 2020) is the construction of contrastive data pairs to facilitate consistency regularization. This involves the incorporation of both strong and weak augmentations, wherein the term *"augmentation"* here means *"distortion"* rather than "enhancement", contrary to intuition. Specifically, strong augmentation ($\mathcal{A}_{strong}$) usually causes significant perturbation to the original data and thus serves as the input for the Student model, while the $\mathcal{A}_{weak}$ produces moderately distorted data samples for the Teacher model to derive better predictions, as demonstrated in the center part of Fig. 2.

❑ **Learning by Labeled *vs.* Unlabeled Data.** In the SSL setting, only a small portion of data is annotated, denoted by $\{x_i, y_i\}_{i=1}^{\mathcal{B}_l}$. The left $\mathcal{B}_u$ samples, $\{x_j\}_{j=1}^{\mathcal{B}_u}$, are all unlabeled. Usually the labeling ratio $\alpha = \frac{\mathcal{B}_l}{\mathcal{B}_l+\mathcal{B}_u}$ is small (*e.g.*, 0.1). Learning based on the labeled data is straightforward by minimizing the cross-entropy loss between model predictions $Pred(x_i)$ and labels y_i:

$$\mathcal{L}_{sup} = \frac{1}{\mathcal{B}_l}\sum_{i=1}^{\mathcal{B}_l}\mathcal{H}(y_i, Pred(x_i)). \tag{2}$$

However, for the unlabeled data x_j, there is no supervision. To solve this, we generate pseudo-labels from the Teacher model predictions $\mathcal{F}^T$, and then calculate the unsupervised

loss as follows:

$$\hat{y_j} = max(\mathcal{F}_t(\mathcal{A}_{weak}(x_j)),$$
$$\mathcal{L}_{un} = \frac{1}{\mathcal{B}_u}\sum_{j=1}^{\mathcal{B}_u} \mathbf{1}(\hat{y_j} > \tau)\mathcal{H}(\hat{y_j}, \mathcal{F}_s(\mathcal{A}_{Strong}(x_j))), \quad (3)$$

where τ is the predefined threshold for confidence scores and $\mathbf{1}$ denotes the indicator function. The whole pipeline is trained using both losses, weighted by hyperparameters,

$$\mathcal{L} = \gamma_1\mathcal{L}_{sup} + \gamma_2\mathcal{L}_{un}. \quad (4)$$

The SeFAR Framework

In this work, we focus on the task of Fine-grained Action Recognition (FAR) in the Semi-Supervised Learning (SSL) setting. This new task brings unprecedented challenges, including: ❶ How to mine abundant and detailed visual information for differentiating subtle differences between fine-grained actions? ❷ How to adapt the original SSL strategies, *e.g.*, consistency regularization, to fit the "temporal-matters" FAR task? ❸ How to deal with the unstable pseudo-labels since the model hesitates between appearance-similar action samples? In the following paragraphs, we will introduce specific designs to address the above challenges.

Dual-level Temporal Elements. Given a fine-grained action video with N frames, we first trim it into K segments (Wang et al. 2016), and randomly sample one frame in each segment, obtaining a frame sequence $\{f_1, ..., f_K\}$ to represent the video. Since in FAR, the high similarity is shared in large parts of visual content (*e.g.*, scenes, objects), models are usually required to perceive subtle changes and abundant details for accurate discrimination. To achieve this, we propose to construct several small temporal elements p_i, where *"small"* means the size L (*i.e.*, the number of containing frames) of p_i is moderate. Intuitively, a small value of L could help the model focus on quick and subtle changes, since details are usually missed when going through too many frames. Given K frames, the sampling step is $\lfloor \frac{K}{L} \rfloor$. After sampling M times, we could get a set of temporal elements with the same temporal lengths, denoted by:

$$\{p_i\}_{i=1}^{M}, \quad |p_i| = L. \quad (5)$$

Besides these temporally fine-grained elements, we also propose to obtain a context element $p^{\textbf{context}}$ to encode long-term information and macro temporal dynamics. $p^{\textbf{context}}$ is composed of more frames, usually two times more than the fine-grained temporal elements p_i. Such dual-level information modeling ensures that multi-granular information is preserved. As a result, we obtain an effective representation of the input video, denoted by $\{p_1, ...p_M, p^{\textbf{context}}\}$.

Perturbation of Fine-grained Temporal Elements. Adopting the FixMatch (Sohn et al. 2020) based semi-supervised learning setting, one key problem is *"how to form the weak-to-strong augmentation pair for consistency regularization"*. For weak augmentation, we could use random horizontal flipping or random scaling, since it largely preserves both spatial and temporal original information. Unfortunately, as pointed out in (Xing et al. 2023), strong augmentation designed for images is insufficient for video tasks, since it fully ignores the temporal dynamics evolving in videos. For the challenging FAR task, temporal variations are even more crucial and require the extreme attention of the model. Therefore, to design more effective strong augmentation strategy $\mathcal{A}_{strong}$ for FAR, we emphasize the following core insights: ① the proposed $\mathcal{A}_{strong}$ should make perturbations to the most crucial part of the data that we want the model to attend to (Sohn et al. 2020; Xie et al. 2020; Kurakin et al. 2020); ② Employing $\mathcal{A}_{strong}$ should not affect the semantic distinctiveness of action categories.

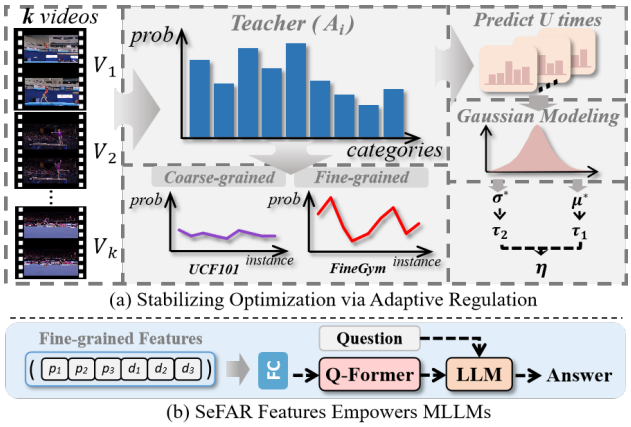

Figure 3: **(a)** For K unlabeled videos, the Teacher model predicts each video multiple times to capture the distribution of predictions, which shows less variability on coarse-grained data and more on fine-grained data. An adaptive coefficient η is calculated from the mean and variance of the distribution to stabilize training. **(b)** MLLM construction pipeline with SeFAR's fine-grained features.

Therefore, combing with the above dual-level temporal modeling strategy, we propose a new strong augmentation operation through introducing temporal perturbation ψ into the fine-grained temporal elements $\{p_i\}$. We experiment with different implementations of ψ, and the final choice is simple but effective: *reversing the frame order*. Specifically, we have:

$$\mathcal{A}_{strong}(\{p_i\}_{i=1}^{M}) = \{\overleftarrow{p_i}\}_{i=1}^{M}, \quad \overleftarrow{p_i} = \psi(p_i) \quad (6)$$

Note that for the temporal context element $p^{\textbf{context}}$, the temporal order is preserved, which ensures the temporal directionality to be inherent in actions (*e.g.*, *"giant circle backward" vs. "giant circle forward"*, etc.), as shown in the bottom-left of Fig. 2. Our augmentation strategy introduces moderate temporal perturbation compared with total shuffling, and it also outperforms previous strategies, *e.g.*, temporal warping (Xing et al. 2023), as shown in Tab. 4.

Stabilizing Optimization via Adaptive Regulation. As mentioned, due to the challenging intrinsic of FAR, models usually swayed precariously between categories with subtle differences. During experiments, the greater the uncertainty of the model's predictions, the less reliable the model's predictions are. Such unstable predictions of the teacher model will result in ambivalent and invalid pseudo-labels for the student, making the whole learning process suffer. To solve this, we first let the *Teacher* model generate predictions U

Method	Backbone	Input	ImgNet	Params	#F	Epoch	Gym99		Gym288		Diving	
							5%	10%	5%	10%	5%	10%
MemDPC (ECCV'20) (Han, Xie, and Zisserman 2020)	3D-ResNet-18	V	✗	15.4M	16	500	10.8	24.1	14.5	21.3	54.3	62.0
LTG (CVPR'22) (Xiao et al. 2022)	3D-ResNet-18	VG	✗	68.3M	8	180	34.3	45.8	16.2	38.7	59.8	64.3
SVFormer (CVPR'23) (Xing et al. 2023)	ViT-B	V	✓	121.4M	8	30	31.4	47.9	21.3	39.6	59.1	70.8
SeFAR-S (Ours)	VIT-S	V	✓	31.2M	8	30	36.7	56.3	27.8	46.9	72.2	78.4
SeFAR-B (Ours)	VIT-B	V	✓	122.1M	8	30	**39.0**	**56.9**	**28.3**	**48.1**	**72.8**	**80.9**

(a) Results of elements across all events.

Method	UB		FX		10m	
	10%	20%	10%	20%	10%	20%
MemDPC	20.7	19.1	13.8	15.9	65.4	71.2
LTG	50.5	60.5	19.6	21.6	75.2	83.5
SVFormer	52.9	66.8	20.1	28.8	73.8	85.9
SeFAR-S (Ours)	56.9	73.8	23.8	42.9	85.5	94.0
SeFAR-B (Ours)	**58.5**	**75.5**	**27.6**	**44.2**	**87.4**	**94.6**

(b) Results of elements within an event.

Method	UB-S1		FX-S1		5253B	
	10%	20%	10%	20%	10%	20%
MemDPC	17.2	21.1	15.4	20.1	82.2	89.5
LTG	21.3	29.7	14.6	19.3	64.6	76.9
SVFormer	28.9	47.3	18.8	22.5	86.6	90.1
SeFAR-S (Ours)	36.6	55.3	19.2	25.5	96.4	97.3
SeFAR-B (Ours)	**37.1**	**56.8**	**20.1**	**26.5**	**97.0**	**97.8**

(c) Results of elements within a set.

Table 1: **Comparison with state-of-the-art semi-supervised action recognition methods on fine-grained datasets.** We employ SeFAR with a sampling combination of {2-2-4}. The primary evaluation metric is top-1 accuracy. In this table, "V" within "Input" denotes RGB video, while "G" represents temporal gradients. "ImgNet" indicates the utilization of models pre-trained on ImageNet (Russakovsky et al. 2015), while "#F" signifies the number of input frames. The labeling rates of the data are indicated by "5%", "10%", and "20%" in the datasets. The best results are highlighted in **Bold**, and the second-best Underlined.

times (U is set to 10 in experiments) for a given unlabeled video, and these predictions may vary largely. Then, based on these, we calculate the mean prediction confidence and standard deviation for each category. For the i^{th} prediction, the predicted probability across all categories constitutes a probability distribution. From this distribution, we can obtain the maximum prediction confidence value μ^i and calculate its standard deviation σ^i. We select the highest confidence value $\mu^* = max(\mu^i)$, along with its corresponding standard deviation σ^*, as shown in Fig. 3.

Based on such μ^* and σ^*, we propose to calculate the dynamic coefficients τ_1 and τ_2 to obtain η, which is further used for adjusting losses derived from unlabeled samples:

$$\begin{aligned} \tau_1 &= sigmoid(e^{\mu^*} - e), \\ \tau_2 &= (sigmoid(\frac{1}{\beta\sigma^* + \epsilon}) - 0.5), \end{aligned} \tag{7}$$

where β is related to the model dropout and ϵ is a steady parameter. To elaborate, τ_1 will increase rapidly as μ^* increases, which enhances high-confidence predictions, while on the other hand, τ_2 suppresses the unstable predictions (*i.e.*, with high standard deviation σ). The obtained adaptive coefficient $\eta = \tau_1 \cdot \tau_2$, is more flexible and beneficial than a predefined hyperparameter. Additionally, for unlabeled data, we also adopt the mixing strategy as in SVFormer (Xing et al. 2023), where the mixture of two unlabeled samples, $\lambda x_1 + (1-\lambda)x_2$, could also serve as input, and the supervision is correspondingly obtained as a mixed version (Details could be found in (Xing et al. 2023)). Here for adjusting $\mathcal{L}_{mix}$, we achieve its coefficient in a similar mixed manner, denoted by $\eta^{'} = \lambda\eta_1 + (1-\lambda)\eta_2$, where η_1, η_2 are individually calculated based on x_1 and x_2. Finally, the total loss of the whole SeFAR framework is as follows:

$$\mathcal{L} = \mathcal{L}_{sup} + \xi(\eta\mathcal{L}_{un} + \eta^{'}\mathcal{L}_{mix}), \tag{8}$$

where $\xi = sin(\frac{n}{M_n})$ is a warmup coefficient calculated using the current epoch number n and the max epoch M_n.

SeFAR Empowers MLLMs. Efforts towards foundation models have led to the development of MLLMs, with vision being the primary modality (Gao et al. 2024). Although shown impressive general capabilities, they may fail in specific and more challenging tasks such as FAR, as illustrated in Fig. 1. This may largely be due to the systematic shortcomings in the visual part as analyzed in (Tong et al. 2024). Given that our SeFAR is designed to be effective for FAR in semi-supervised scenarios, the question: *"Could SeFAR benefit current MLLMs through providing better visual perception?"* The answer is yes as supported by the results in Tab. 5. To elaborate, in line with the typical MLLM framework, a frozen visual encoder is usually combined with a LLM. This setup facilitates multimodal functionality by aligning visual and textual features using an adaptor, *e.g.*, Q-Former (Li et al. 2023a). Given such a setting, we could use the features extracted by SeFAR to replace those provided by the original visual encoder as shown at the bottom of Fig. 3. Similarly, by aligning the visual features with the textual domain and concatenating with text embeddings, we could feed them into the LLM to produce the answers. Results show that SeFAR features could lead to much better results compared to those used in original MLLM settings.

Experiment

Experiment Setup

Datasets and Evaluation. We perform evaluations on fine-grained datasets Gym99, Gym288 (Shao et al. 2020), and FineDiving (Xu et al. 2022a), as well as coarse-grained datasets UCF-101 (Soomro 2012) and HMDB-51 (Kuehne et al. 2011), using Top-1 accuracy as metrics. Specifically, FineGym includes hierarchical annotations at three semantic granularity: *events*, *sets*, and *elements*. All the experiments on FineGym are performed at the element level, but within different scopes. FineDiving is a diving dataset comprising 3000 annotated clips with timestamps, encompassing 52 *ac-*

Method	Backbone	Input	ImgNet	#F	Epoch	UCF-101			HMDB-51	
						1%	5%	10%	40%	50%
MT+SD (WACV'21) (Jing et al. 2021)	3D-ResNet-18	V	✗	16	500	-	31.2	40.7	32.6	35.1
MvPL (ICCV'21) (Xiong et al. 2021)	3D-ResNet-50	VFG	✗	8	600	22.8	41.2	80.5	30.5	33.9
TCLR (CVIU'22) (Dave et al. 2022)	3D-ResNet-18	V	✗	16	1200	26.9	-	66.1	-	-
CMPL (CVPR'22) (Xu et al. 2022b)	R50+R50-1/4	V	✗	8	200	25.1	-	79.1	-	-
LTG (CVPR'22) (Xiao et al. 2022)	3D-ResNet-18	VG	✗	8	180	-	44.8	62.4	46.5	48.4
TimeBalance (CVPR'23) (Dave et al. 2023)	3D-ResNet-50	V	✗	8	250	30.1	53.3	81.1	52.6	53.9
SeFAR (Ours)	VIT-S	V	✗	8	30	35.2	64.1	78.3	55.9	59.2
FixMatch (NeurIPS'20) (Sohn et al. 2020)	SlowFast-R50	V	✓	8	200	16.1	-	55.1	-	-
MemDPC (ECCV'20) (Han, Xie, and Zisserman 2020)	3D-ResNet-18	V	✓	16	500	-	-	44.2	-	-
ActorCM (CVIU'21) (Zou et al. 2023)	R(2+1)D-34	V	✓	8	360	-	45.1	53.0	35.7	39.5
VideoSSL (WACV'21) (Jing et al. 2021)	3D-ResNet-18	V	✓	16	500	-	32.4	42.0	32.7	36.2
TACL (TSVT'22) (Tong, Tang, and Wang 2023)	3D-ResNet-50	V	✓	16	200	-	35.6	55.6	38.7	40.2
L2A (ECCV'22) (Gowda et al. 2022)	3D-ResNet-18	V	✓	8	400	-	-	60.1	42.1	46.3
SVFormer-S (CVPR'23) (Xing et al. 2023)	ViT-S	V	✓	8	30	31.4	-	79.1	56.2	58.2
SVFormer-B (CVPR'23) (Xing et al. 2023)	ViT-B	V	✓	8	30	46.1	-	84.6	59.9	64.3
SeFAR (Ours)	VIT-S	V	✓	8	30	46.0	73.2	84.3	58.5	62.9
SeFAR (Ours)	VIT-B	V	✓	8	30	**50.3**	**77.6**	**87.0**	**61.5**	**65.7**

Table 2: **Comparison with state-of-the-art semi-supervised action recognition methods on coarse-grained datasets.**"V" within "Input" signifies RGB video, "F" indicates optical flow, while "G" denotes temporal gradients.

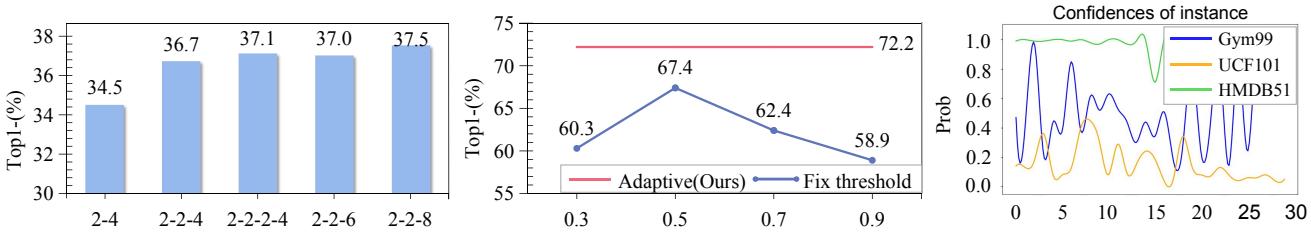

Figure 4: **Ablation Studies.** We compare SeFAR-B with different sampling combinations on Gym-99 5%, as illustrated on the **left**. We also contrast fixed threshold methods with our Adaptive Regulation strategy on FineDiving 5% in the **middle**. On the **right** side, we demonstrate the fluctuation of predictions made by the Teacher model across different datasets.

tion types, 29 *sub-action* types, and 23 difficulty levels.

Baselines. We employ the ViT (Dosovitskiy 2020) extended model TimeSformer (Bertasius, Wang, and Torresani 2021) as the backbone. We instantiate the SeFAR-S model based on ViT-S, with the number of total parameters comparable to most previous Conv-based methods (Han, Xie, and Zisserman 2020; Xiong et al. 2021; Xu et al. 2022b; Xiao et al. 2022; Tong, Tang, and Wang 2023; Gowda et al. 2022; Dave et al. 2023). Moreover, we implement the SeFAR-B model based on ViT-B, with more parameters. We configure the sampling combination by default as $\{2-2-4\}$ for SeFAR, as commonly used 8-frame input.

Main Results

The main quantitative results on the two fine-grained action recognition datasets, *i.e.*, FineGym and FineDiving, are demonstrated in Tab. 1. We evaluate all the methods at different semantic granularities. Specifically, we first conduct experiments on Gym99 and Gym288. Then, by narrowing the semantic scope, we focus on those element-level categories belonging to a specific event. For instance, in Gym99, 25 classes belong to Uneven-bars (UB), while 35 classes are from Floor-exercise (FX). Further, we delve into the finer granularity, collecting sampling within that same set in the same event. Here we get all the circles in UB-set1 (UB-S1) and all the jumps in FX-set1 (FX-S1) for evaluation. We can observe that on both the FineGym and FineDiving, SeFAR-S significantly outperforms previous *open-sourced* semi-supervised action recognition methods across all semantic granularities with moderate parameters. Additionally, when increasing the parameters comparative with SVFormer (Xing et al. 2023), the larger model, SeFAR-B, performs even better. Both SeFAR-S and SeFAR-B display the effectiveness of our proposed SeFAR framework for addressing the challenging task.

Moreover, to further inspect the effectiveness and robustness of SeFAR, we conducted experiments on two classical coarse-grained action recognition datasets, UCF-101 and HMDB-51. As shown in Tab. 2, SeFAR-B achieves approximately 3.3% improvement on UCF101 and approximately 1.7% improvement on HMDB51, achieving new state-of-the-art results compared with those competitive baselines.

Ablation Studies

To achieve an in-depth comprehension of our SeFAR framework, we perform ablation studies on the impact of each component, namely *dual-level temporal elements modeling* (Dual-Ele), *moderate temporal perturbation* (Mod-Perturb) and *Adaptive Regulation* (Ada-Reg), as demonstrated in Tab. 3. Each module contributes significantly as an essential part of SeFAR. Details can be found as follows.

Analysis of Dual-level Temporal Elements Modeling. We first compare different combinations of sampled elements, each context element has varying temporal lengths,

Dual-Ele	Mod-Perturb	Ada-Reg	Gym99	Gym288	Diving
✗	✗	✗	32.6	22.7	60.4
✓	✗	✗	34.8	25.4	64.6
✓	✓	✗	35.9	26.6	67.4
✓	✓	✓	**36.7**	**27.8**	**72.2**

Table 3: **Ablations of different components with SeFAR**, where ✓ means "w/". We employ temporal warping consistent with SVFormer once our Mod-Perturb is eliminated.

Perturbation	S/O	Gym99	Gym288	Diving	G.-New	Sth.-Sth.
Spatial-only		34.2	24.4	67.9	45.6	39.4
Slow (T-Drop)	S	35.6	25.2	68.6	45.0	41.2
All shuffle	O	35.2	26.3	69.0	45.5	41.9
Local-shuffle	O	36.4	27.6	71.9	45.3	43.3
Warping	O	35.9	24.7	68.2	44.8	40.8
T-Half	O	36.0	24.8	68.4	44.8	42.1
All reverse	O	36.3	27.3	71.2	45.9	42.7
Mod-Perturb	O	**36.7**	**27.8**	**72.2**	**46.2**	**44.9**

Table 4: **Ablation of different temporal augmentations. S** and **O** denote the Speed- and Order-focused.

e.g., $4, 6, 8$. To facilitate comparison, we fix the length of the temporal fine-grained elements to be 2, consistent with our default setting $\{2\text{-}2\text{-}4\}$. Results are depicted in the left part of Fig. 4. We can find that even with a limited input of only 6 frames, *i.e.*, $\{2\text{-}4\}$, our proposed SeFAR surpasses the 8-frame input baseline SVFormer (Xing et al. 2023). This observation justifies the capability of our *dual-level temporal elements modeling* to capture abundant information details from video data, contributing to better discerning subtle differences among fine-grained actions. Additionally, it is noteworthy that increasing the number of the fine-grained elements, *i.e.*, $\{2\text{-}2\text{-}2\text{-}4\}$, or extending the temporal length of the context element, *i.e.*, $\{2\text{-}2\text{-}6\}$ and $\{2\text{-}2\text{-}8\}$, all leads to performance improvements. This is attributed to the fact that more frames entail richer action information.

Analysis of Moderate Temporal Perturbation. To better explore the impact of our proposed moderate temporal perturbation (Mod-Perturb), we first selected 40 classes of action pairs that are reversing to each other (*e.g.*, "giant circle backward" *vs.* "giant circle forward") from FineGym, forming a subset called **Gym-New** (G.-New). As shown in Tab. 4, SeFAR also maintains superior performance even on such actions, as well as on the Something-Something V2 (Sth.-Sth.) dataset (Goyal et al. 2017). Furthermore, we compare our Mod-Perturb with other temporal perturbation strategies in both Speed- and Order-focused (*e.g.*, slow-rate (Singh et al. 2021), temporal warping (Xing et al. 2023), T-Drop and T-Half (Zou et al. 2023)), the results can be found in Tab. 4. We can observe that: 1) Our Mod-Perturb exhibits superior stability and efficacy compared to other temporal augmentations and spatial-only (temporal information well-kept). 2) Spatial-only is less effective in Gym99 but outperforms most temporally augmented in Gym-New. This suggests that preserving accurate temporal information is crucial for more complex datasets, whereas reasonable temporal perturbations can enhance model stability in larger and more diverse datasets, and Mod-Perturb benefits from both.

Analysis of Adaptive Regulation. To justify the usefulness of our stabilizing coefficients for adaptive losses, we perform two analyses: ① We compare this strategy with the fixed thresholding strategy widely used in the classical SSL method, the results are displayed in Fig. 4 (b), showing our method is both stable and effective. ② In Fig. 4 (c), We demonstrate the unstable predictions provided by the teacher models for FAR. Specifically, we randomly draw 30 data samples from different datasets, UCF101, HMDB51, and FineGym, for the teacher model to offer predictions. The highly varying predictions on FineGym further justify the motivation of our stabilizing design for FAR.

Visual Encoder	MLLM	Gym-QA-99	Gym-QA-288
CLIP-ViT-L/16	,	37.3	41.0
EVA-CLIP ViT-G/14	,	43.7	44.8
ViT-L/14		44.3	46.0
SeFAR (Ours)	-	**49.0**	**56.2**

Table 5: **Ablation of Pre-trained Visual Encoder.** We employ Vicuna-7B (Chiang et al. 2023) as the base LLM, comparing SeFAR's features with the pre-trained features of commonly used visual encoders in MLLMs further fine-tuned on 5% data (*i.e.*, : LLaVA, : VideoChat2, : VideoLLaMA, : VideoChat, and : VideoLLaVA)

Analysis of SeFAR Features. To further demonstrate the capability of SeFAR in enhancing MLLMs, we first constructed the **Gym-QA** dataset, which is derived from FineGym and presented in a multiple-choice format as illustrated in Fig. 1. We then selected three widely used MLLM visual encoders, *i.e.*, CLIP-ViT-L/16, EVA-CLIP ViT-G/14, and ViT-L/14). For fair comparisons, we conduct semi-supervised training on these backbones with 5% labeling data from FineGym. Subsequently, we froze the weights of these encoders along with the weights from our 5%-trained SeFAR, and fine-tuned the Q-former using 5% of the annotated data from Gym-QA. As shown in Tab. 5, the SeFAR-empowered LLM significantly outperformed the other MLLM visual encoders on the Gym-QA task. This also mitigates the challenge of fine-tuning MLLMs in scenarios with low labeling rates.

Conclusion

In this work, we shed light on a more challenging and specific video understanding task, Semi-supervised Fine-grained Action Recognition (FAR). To tackle this, we propose SeFAR, which adopts ideas from the FixMatch setting and possesses innovative components delicately devised for FAR. Specifically, SeFAR is distinguished due to the following designs: 1) *Dual-level temporal elements modeling* is used to mine visual cues more thoroughly and capture rich temporal dynamics better; 2) *Augmentation via moderate temporal perturbation* is to produce temporally strong-distorted samples for weak-to-strong consistency regularization; 3) *Stabilizing Optimization via Adaptive Regulation* is to address the large uncertainty in model predictions. To highlight, SeFAR also demonstrates superior performance in empowering MLLM's fine-grained visual understanding capability. SeFAR outperforms all the baselines largely on representative both fine- and coarse-grained datasets.

Acknowledgments

This work was founded by the National Natural Science Foundation of China (NSFC) under Grant 62306239, and was also supported by National Key Lab of Unmanned Aerial Vehicle Technology under Grant WR202413.

References

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Dave, I.; Gupta, R.; Rizve, M. N.; and Shah, M. 2022. TCLR: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219: 103406.

Dave, I. R.; Rizve, M. N.; Chen, C.; and Shah, M. 2023. Timebalance: Temporally-invariant and temporally-distinctive video representations for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2341–2352.

DeVries, T. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Gao, P.; Zhang, R.; Liu, C.; Qiu, L.; Huang, S.; Lin, W.; Zhao, S.; Geng, S.; Lin, Z.; Jin, P.; et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Gowda, S. N.; Rohrbach, M.; Keller, F.; and Sevilla-Lara, L. 2022. Learn2augment: learning to composite videos for data augmentation in action recognition. In *European conference on computer vision*, 242–259. Springer.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, 5842–5850.

Han, T.; Xie, W.; and Zisserman, A. 2020. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, 312–329. Springer.

Hu, A.; Shi, Y.; Xu, H.; Ye, J.; Ye, Q.; Yan, M.; Li, C.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6929–6938.

Jing, L.; Parag, T.; Wu, Z.; Tian, Y.; and Wang, H. 2021. Videossl: Semi-supervised learning for video classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1110–1119.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, 2556–2563.

Kurakin, A.; Raffel, C.; Berthelot, D.; Cubuk, E. D.; Zhang, H.; Sohn, K.; and Carlini, N. 2020. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring.

Leong, M. C.; Zhang, H.; Tan, H. L.; Li, L.; and Lim, J. H. 2022. Combined CNN transformer encoder for enhanced fine-grained human action recognition. *arXiv preprint arXiv:2208.01897*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, T.; Foo, L. G.; Ke, Q.; Rahmani, H.; Wang, A.; Wang, J.; and Liu, J. 2022. Dynamic spatio-temporal specialization learning for fine-grained action recognition. In *European Conference on Computer Vision*, 386–403. Springer.

Li, Z.; He, L.; and Xu, H. 2022. Weakly-supervised temporal action detection for fine-grained videos with hierarchical atomic actions. In *European conference on computer vision*, 567–584. Springer.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Mac, K.-N. C.; Joshi, D.; Yeh, R. A.; Xiong, J.; Feris, R. S.; and Do, M. N. 2019. Learning motion in feature space: Locally-consistent deformable convolution networks for fine-grained action detection. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, 6282–6291.

OpenAI. 2024. GPT-4 System Card. https://openai.com/index/gpt-4v-system-card/. Accessed: 2024-08-03.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Shao, D.; Zhao, Y.; Dai, B.; and Lin, D. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2616–2625.

Singh, A.; Chakraborty, O.; Varshney, A.; Panda, R.; Feris, R.; Saenko, K.; and Das, A. 2021. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10389–10399.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Soomro, K. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *Proceedings of the 31st ACM international conference on multimedia*, 1719–1728.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tong, A.; Tang, C.; and Wang, W. 2023. Semi-Supervised Action Recognition From Temporal Augmentation Using Curriculum Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3): 1305–1319.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9568–9578.

Vemprala, S. H.; Bonatti, R.; Bucker, A.; and Kapoor, A. 2024. Chatgpt for robotics: Design principles and model abilities. *IEEE Access*.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, 20–36. Springer.

Xiao, J.; Jing, L.; Zhang, L.; He, J.; She, Q.; Zhou, Z.; Yuille, A.; and Li, Y. 2022. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3252–3262.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33: 6256–6268.

Xing, Z.; Dai, Q.; Hu, H.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2023. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18816–18826.

Xiong, B.; Fan, H.; Grauman, K.; and Feichtenhofer, C. 2021. Multiview pseudo-labeling for semi-supervised learning from video. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7209–7219.

Xu, J.; Rao, Y.; Yu, X.; Chen, G.; Zhou, J.; and Lu, J. 2022a. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2949–2958.

Xu, Y.; Wei, F.; Sun, X.; Yang, C.; Shen, Y.; Dai, B.; Zhou, B.; and Lin, S. 2022b. Cross-model pseudo-labeling for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2959–2968.

Yang, C.; Xu, Y.; Shi, J.; Dai, B.; and Zhou, B. 2020. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 591–600.

Yuan, L.; Gundavarapu, N. B.; Zhao, L.; Zhou, H.; Cui, Y.; Jiang, L.; Yang, X.; Jia, M.; Weyand, T.; Friedman, L.; et al. 2023. Videoglue: Video general understanding evaluation of foundation models. *arXiv preprint arXiv:2307.03166*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, P.; Dong, X.; Zang, Y.; Cao, Y.; Qian, R.; Chen, L.; Guo, Q.; Duan, H.; Wang, B.; Ouyang, L.; Zhang, S.; Zhang, W.; Li, Y.; Gao, Y.; Sun, P.; Zhang, X.; Li, W.; Li, J.; Wang, W.; Yan, H.; He, C.; Zhang, X.; Chen, K.; Dai, J.; Qiao, Y.; Lin, D.; and Wang, J. 2024. InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output. *arXiv preprint arXiv:2407.03320*.

Zhao, L.; Gundavarapu, N. B.; Yuan, L.; Zhou, H.; Yan, S.; Sun, J. J.; Friedman, L.; Qian, R.; Weyand, T.; Zhao, Y.; et al. 2024. VideoPrism: A Foundational Visual Encoder for Video Understanding. *arXiv preprint arXiv:2402.13217*.

Zhou, B.; Andonian, A.; Oliva, A.; and Torralba, A. 2018. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, 803–818.

Zhu, X. J. 2005. Semi-supervised learning literature survey.

Zou, Y.; Choi, J.; Wang, Q.; and Huang, J.-B. 2023. Learning representational invariances for data-efficient action recognition. *Computer Vision and Image Understanding*, 227: 103597.