Supplementary: FinePseudo: Improving Pseudo-Labelling through Temporal-Alignablity for Semi-Supervised Fine-Grained Action Recognition

Ishan Rajendrakumar Dave, Mamshad Nayeem Rizve, and Mubarak Shah

Center for Research in Computer Vision, University of Central Florida, USA
ishandave@ucf.edu, nayeemrizve@gmail.com, shah@crcv.ucf.edu
https://daveishan.github.io/finepsuedo-webpage/

Overview

- Section A: Details of the datasets
- Section B: Implementation details about the hyperparameters and training schedule
- Section C: Additional ablation for our framework
- Section D: Results on additional splits and tasks.
- Section E: Supportive algorithm and diagrams
- Section F: Detailed comparison with related prior work

A Dataset Details

All datasets used in our study are publicly available. We utilize only the action class labels from these datasets.

Diving48 [17] includes 48 action classes of diving actions. Each sequence is defined by a combination of takeoff (dive groups), movements in flight (somersaults and/or twists), and entry (dive positions). We utilize the V2 set of annotations, which is a cleaner version.

FineGym [22] provides challenging fine-grained action classes of various gymnastic events. Some samples from this dataset are shown in Fig. 1. Apart from FineGym99 and FineGym288 mentioned in the main paper, we also present results within each of the event subsets, as used in recent work [24].

Vault (VT) [22] contains 6 action classes from the Vault event. Its training/test split contains 1k/0.5k videos.

Floor (FX) [22] includes 35 action classes from the 'Floor Exercise' event. Its training/test split contains 5.3k/2.2k videos.

UB-S1 [22] comprises 15 action classes covering videos of different types of circles around the bars. Its training/test split contains 3.5k/1.5k videos.

FX-S1 [22] is a subset of the Floor Exercise (FX) set, covering 11 actions related to leaps, jumps, and hops. Its training/test split contains 1.9k/0.7k videos.

FineDiving [29] includes approximately 3k videos covering 52 action classes from Diving sequences. This dataset focuses on the problem of action quality assessment, providing annotations for steps and scores. However, we utilize only the 'action' annotations in our work.
Kinetics400 [3] contains more general human actions collected from YouTube. It covers 400 action classes, with a training/validation split of 240k/20k videos.
Something-Something V2 [12] focuses on actions related to hand-object interactions. We utilize a split from prior work [23, 33], which covers 82k training and 12k test videos.

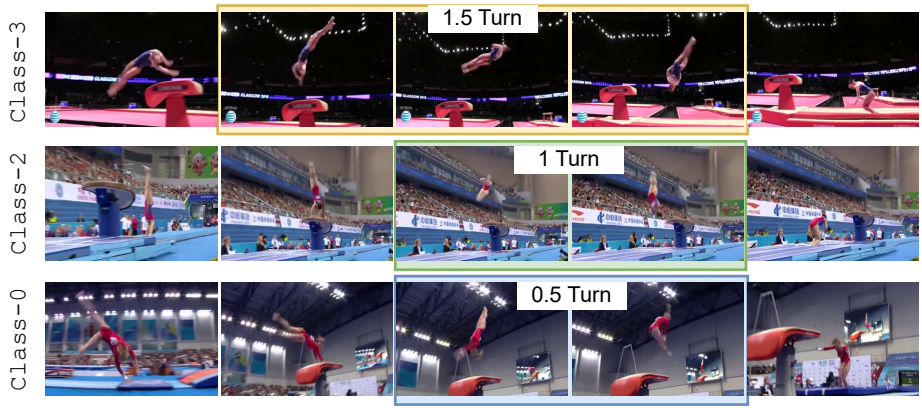

Fig. 1: Samples from the FineGym Dataset. FineGym offers a range of challenging, fine-grained action classes derived from gymnastic events. This figure showcases three action classes from the FineGym288 split. Here, each action class differs in the phase where different numbers of turns are executed.

B Implementation Details

Network Architecture: Alignability Encoder (f_A) is a Video Transformer Network (VTN) [19] architecture following prior work [5,31]. For non-linear project head $g(\cdot)$ we employ a multilayer perception (MLP) following [6]. For Action Encoder (f_E) we utilize the R2plus1D-18 [25] model by default, which is initialized with SSL pretraining [8] on the given dataset. For the score mapping function f_S we utilize a 2-layer MLP.
Training: SSL pretraining of f_A takes place for 100 epochs. Alignability-verification based metric learning of f_A and training of f_E takes 100 epochs. In the self-training steps, the proposed collaborative PL generation each takes place at every 5th epoch of labeled training, this process runs for 10 training iterations.
Inference For inference, we only consider the video encoder f_E, following a commonly used protocol [25]. We first obtain clip-level predictions from 10 uni-

formly sampled clips across the video duration and 3 spatial crops, then average these predictions to derive a video-level prediction.

B.1 Hyperparameters

SSL pretraining of f_A For the Gaussian Infused Temporal Distinctiveness Loss ($\mathcal{L}_{GITDL}$) (Eq. 1), we set the temperature parameter (τ) to 0.1. Additionally, for the Gaussian prior, we use a peak value (κ) of 0.99 and a standard deviation (σ) of 0.2.

Alignability-based Metric Learning After SSL pretraining of f_A, we freeze the image encoder and continue training only the temporal encoder of the VTN architecture. For the computation of softDTW, we set the smoothness parameter (γ) to 0.001. In the case of the Alignability-based Triplet Loss ($\mathcal{L}_{AT}$), we use a default margin (m) of 0.1. Our batch size is set to 96, and we employ a subsampler in the dataloader to ensure that there are at least two instances from each sampled action class.

Collaborative Pseudolabeling process To construct the embedding set $\mathbb{A}$ from the labeled dataset, we randomly select $\rho = \min(15, \text{samples in the class})$ samples from each class. For the non-parametric classifier (as detailed in Eq. 8 of the main paper), we set the temperature parameter τ to 0.1. The confidence threshold θ is established at 0.6.

For our collaborative pseudo-labeling process, only a single forward pass is sufficient for each video in both $\mathbb{D}_l$ and $\mathbb{D}_u$ to extract their respective features. Subsequently, the classwise alignability score is computed in parallel on these extracted features, significantly enhancing the speed of the pseudo-labeling process and not bottlenecking the speed of the overall PL process.

B.2 Optimization and Training Schedule

To update the parameters of the network, we employ the Adam optimizer [16], using its default parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the learning rate scheduler, we apply a base learning rate of 10^{-4}, accompanied by a linear warmup over the first 5 epochs, followed by a cosine decay learning rate scheduler.

C Additional Ablations

For additional ablation, we follow the same default setup of the ablation of the main paper *i.e.* reporting results on the action recognition task of various fractions of labeled set of Diving48 dataset with R2plus1D model.

C.1 Ablation with Triplet Mining Strategies

For our mini-batch sampling, we ensure that each class sampled has at least two instances. We then calculate the alignment cost (as per Eq. 1 in the main paper) between each pair of samples within the mini-batch. Samples from the same class serve as positives, while pairs from different classes are considered negatives. While all positive pairs are included in our analysis, we explore various strategies for mining negative pairs in Table 1.

Table 1: Ablation with Triplet loss

Triplet Mining	**10%**	**20%**
All Negatives	36.16	58.65
Hard Negatives	**37.64**	**60.40**
Hardest Negative only	37.20	**60.40**

In the first row, where all negative pairs are considered, we observe less effective learning. This is due to easy negatives (where $D^n - D^p < m$) that fail to effectively contribute significantly to the learning process. On the other hand, mining hard negatives—specifically, considering only those negative pairs where $D^n - D^p > m$—and selecting the hardest negative from the mini-batch, shows improved performance. However, the 'hardest-negative' strategy performs slightly worse than the 'hard negatives' in the 10% data scenario, likely due to the reduced number of available triplets.

C.2 Empirical evidence: suitability of Alignment based distance

We conduct experiments utilizing various distance functions D in Eq. 2 of the main paper to train f_A using our proposed metric learning approach. Given that our metric learning is centered around a verification task (determining whether a video pair belongs to the same class), we also report the validation average precision in Table 2. The findings reveal that the alignment cost(softDTW) markedly outperforms other distance measures across diverse tasks. Moreover, for fine-grained action categories, a distance function based on alignment is far more effective than the standard cosine distance, underscoring our motivation Fig. 1(c) of the main paper.

SSL pretraining of Alignability-encoder f_A:

To assess the representation quality of SSL pretraining of f_A (Supp. Sec. E), we conduct additional evaluations on fine-grained video tasks of the PennAction dataset [32]: Phase Classification and Event Progress, following the protocol in [11]. These tasks are action-class agnostic and require an understanding of the action phase.

Results from Table 3 suggest that our proposed Gaussian prior-based frame-level temporal distinctiveness significantly improves the performance of phase-level tasks and the overall video-level semi-supervised learning performance on

Table 2: Ablation of different distance in metric learning

Distance Type	AP	10%	20%
cosine- mean	0.57	33.41	53.60
cosine- full seq.	0.48	32.15	52.54
cosine- 4 seg	0.64	34.90	54.51
OTAM [2]	0.68	35.06	56.33
softDTW [7]	**0.72**	**37.64**	**60.40**

Table 3: Ablation: SSL pretraining of f_A

SSL Objective	PennAction		Diving48	
	Phase Classi.	Event Progress	10%	20%
w/o gaussian	0.88	0.87	35.42	58.81
with gaussian	**0.93**	**0.91**	**37.64**	**60.40**

fine-grained actions. This improvement is attributed to the Gaussian prior, which enhances temporal coherence (smoothness) in the frame-wise video embedding.

D Additional Results

D.1 Results with ImageNet Pretraining

We additionally present results using the ViT-B backbone, pretrained on ImageNet [10], and apply it to both fine-grained (Diving48) and coarse-grained (Kinetics400) datasets. These results are presented in Table 4. Our method surpasses the performance of the previous approach [27], which employs the same backbone and pretrained weights.

Table 4: Results with backbone initialization from ImageNet (supervised)

Method	Backbone	Diving48		Kinetics400	
		10%	20%	1%	10%
SVFormer-B [27]	ViT-B	49.7	71.1	49.1	69.4
Ours	ViT-B	**54.2**	**75.7**	**52.0**	**71.1**

D.2 Results on FineGym subsets

Results on the Standard FineGym99/288 splits, which encompass all four types of gymnastic events—Vault, Floor Exercise, Balance Beam, and Uneven Bars—are

Table 5: Comparison with prior work of fine-grained video understanding on Action Recognition task.

Method	% labels	Model	Init. Data	FG99	FG288
D^3TW CVPR'19 [4]	100%	R(2D+3D)-50	Labeled ImageNet	15.3	14.1
SpeedNet CVPR'20 [1]	100%	R(2D+3D)-50	Labeled ImageNet	16.9	15.6
TCN ICRA'18 [21]	100%	R(2D+3D)-50	Labeled ImageNet	20.0	17.1
SaL ECCV'16 [18]	100%	R(2D+3D)-50	Labeled ImageNet	21.5	19.6
TCC CVPR'19 [11]	100%	R(2D+3D)-50	Labeled ImageNet	25.2	20.8
GTA CVPR'21 [13]	100%	R(2D+3D)-50	Labeled ImageNet	27.8	24.2
CARL CVPR'22 [5]	100%	VTN (R50)	Unlabeled ImageNet	41.8	35.2
VSP CVPR'23 [31]	100%	VTN (R50)	Unlabeled ImageNet	43.1	36.9
VSP-P CVPR'23 [31]	100%	VTN (R50)	Unlabeled ImageNet	44.6	38.2
VSP-F CVPR'23 [31]	100%	VTN (R50)	Unlabeled ImageNet	45.7	39.5
Ours(*FinePseudo*)	5%	VTN (R50)	Unlabeled ImageNet	41.1	34.4
Ours(*FinePseudo*)	10%	VTN (R50)	Unlabeled ImageNet	**66.2**	**56.5**

presented. The action classes from these diverse events are semantically distinct from one another. In our analysis, we treat actions from each event separately, adding further complexity to the classification problem. The results are detailed in Table 6. Initially, we evaluate video self-supervised learning baselines: TCLR [8] and VideoMoCo [20]. Subsequently, models initialized with the weights from [8] are used to assess semi-supervised methods. Our method consistently outperforms previous methods by a significant margin across all splits. This indicates the superior ability of our semi-supervised approach to distinguish fine-grained, semantically similar actions within each event set.

Table 6: Results on within set activities of FineGym dataset

Method	Vault (VT)			Floor (FX)			UB-S1			FX-S1		
	5%	10%	20%	5%	10%	20%	5%	10%	20%	5%	10%	20%
TCLR CVIU'22 [8]	34.2	39.7	41.6	24.0	25.4	57.6	22.5	41.7	60.6	17.9	21.6	34.8
VidMoCo CVPR'21 [20]	32.0	38.9	40.7	22.3	23.6	55.1	19.8	40.3	59.2	14.6	18.9	32.5
PL	34.1	39.9	42.4	23.9	25.7	58.1	22.8	42.3	62.5	17.4	21.5	35.1
TimeBal CVPR'23 [9]	35.7	40.4	43.1	24.6	26.3	59.7	28.6	43.1	63.2	19.2	22.3	35.5
Ours(*FinePseudo*)	**40.8**	**44.0**	**47.6**	**29.2**	**30.0**	**63.6**	**32.4**	**46.5**	**67.4**	**23.5**	**27.7**	**39.2**

D.3 Comparison with fine-grained video methods

Additionally, we compare our results with previous methods that specialize in video fine-grained intra-video tasks, as shown in Table 5. Without the need for extra data, our method surpasses these prior approaches by leveraging only 10% of the labeled data.

D.4 Results on Class-agnostic Fine-grained tasks

While our primary focus is on semi-supervised action recognition, we also present the performance of our alignability encoder f_A on class-agnostic fine-grained tasks such as Phase Classification, Kendall's Tau, and Event Progress, as proposed by [11]. We evaluate f_A directly following SSL pretraining, without the use of any labeled data. The results, detailed in Table 7, demonstrate that our method performs favorably compared to those specialized in these tasks. It also shows the effectiveness of our GITDL-based SSL pretraining in capturing tasks that are based on intra-video dynamics, such as action-phases.

Table 7: Results on fine-grained tasks of PennAction dataset [32].

Method	Label Used	Phase Classi.	Kendall's Tau	Event Progress
TCC CVPR'19 [11]	Action	0.744	0.641	0.591
GTA CVPR'21 [13]	Action	-	0.748	-
LAV CVPR'21 [14]	Action	0.786	0.684	0.625
SaL ECCV'16 [18]	None	0.682	0.474	0.390
TCN ICRA'18 [21]	None	0.681	0.542	0.383
CARL CVPR'22 [5]	None	0.931	0.985	0.918
VSP CVPR'23 [31]	None	0.931	0.986	**0.923**
Ours (f_A)	None	**0.932**	**0.992**	0.911

D.5 Complementary behavior- VideoSSL methods

To substantiate the claims, we analyze two distinct types of video SSL methods: (1) TCLR, which focuses on learning video-level representations for high-level semantic tasks such as action recognition, and (2) CARL, oriented towards learning frame-level video representations for low-level intra-video tasks like phase classification.

In our analysis, we utilize publicly available Kinetics400 pre-trained weights for both TCLR and CARL. We then evaluate their performance on intra-video tasks using the PennAction dataset [32] and on the video-level action recognition task with the Diving48 dataset [17], as detailed in Table 8. This comparison reveals distinct behavioral patterns of the two video SSL methods across these tasks.

E Method

E.1 SSL pretraining of Alignability encoder

Given the limited scale of labeled data ($\mathbb{D}_l$), our primary objective is to effectively utilize the extensive scale of unlabeled data ($\mathbb{D}_u$) to facilitate the learning of

Table 8: Complementary behavior of VideoSSL methods.

Method	PennAction		Diving48	
	Phase Class.	Kendall's Tau	10%	20%
CARL [5]	0.931	0.985	26.8	47.1
TCLR [8]	0.799	0.821	33.1	53.7

frame-wise video representations in f_A, which can useful to identify the action phase.

Recent advancements in clip-level video self-supervised methods, have shown promising results in learning powerful representations within a single video instance by employing a temporal-distinctiveness objective [8, 9]. In the standard *clip-level* temporal-distinctiveness formulation, within a video instance, temporally-aligned clips are treated as positive, while temporally-misaligned clips are considered negative. However, this approach treats each misaligned timestamp equally negative, regardless of their temporal distance from the anchor clip. In the context of *frame-level* video representations, treating negative equally loses the frame-wise temporal coherence (smoothness). As established by prior work [5, 11, 13, 14, 31], it is crucial for capturing intra-video dynamics, such as action phases. To address this and achieve temporal coherence in learning frame-wise temporal-distinctiveness, we introduce a gaussian kernel to the negative timestamps. This modification ensures that the weight of a negative instance increases smoothly (due to gaussian) and proportionally with its timestamp difference from the anchor.

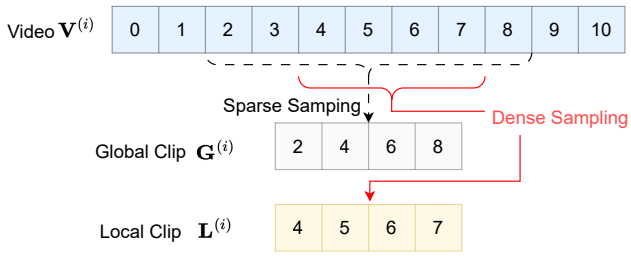

Fig. 2: Clip Sampling in the Proposed GITDL Framework. From a full video $\mathbf{V}^{(i)}$, we sample two types of clips: a global clip $\mathbf{G}^{(i)}$, which is sparsely sampled (skip rate = 2), and a local clip $\mathbf{L}^{(i)}$, which is densely sampled (skip rate = 1) within the temporal range of $\mathbf{G}^{(i)}$.

Consider a video instance i from which we sample a global clip G and a local clip L, with L being a subset of G. Both clips are sampled to have exactly T frames - G through sparse sampling and L through dense sampling (Visual Aid in Fig. 2). These clips are then fed into the alignability encoder f_A and a non-linear projection layer $g(\cdot)$, resulting in their frame-wise video representations

$\{\bar{\mathbf{g}}_t^i\}_{t=1}^{T}$ and $\{\bar{\mathbf{l}}_t^i\}_{t=1}^{T}$. Next, we subsample these representations to retain only the frame-ids present in both clips. This results in temporally corresponding representations with $\mathcal{T}$ frames $\{\mathbf{g}_t^i\}_{t=1}^{\mathcal{T}}$ and $\{\mathbf{l}_t^i\}_{t=1}^{\mathcal{T}}$. Our novel objective, Gaussian Infused Temporal Distinctiveness Learning(GITDL), is formulated as follows:

$$\mathcal{L}_{GITDL}^{(i)} = -\sum_{t_1=1}^{\mathcal{T}} \log \frac{\mathrm{h}(\mathbf{l}_{t_1}^{(i)}, \mathbf{g}_{t_1}^{(i)})}{\sum_{\substack{t_2=1 \\ t_2 \neq t_1}}^{\mathcal{T}} (1 - \kappa e^{-\frac{(t_1-t_2)^2}{2\sigma^2}}) \mathrm{h}(\mathbf{l}_{t_1}^{(i)}, \mathbf{g}_{t_2}^{(i)})} \tag{1}$$

Where $\mathrm{h}(\mathbf{u_1}, \mathbf{u_2}) = \exp\left(\frac{\mathbf{u_1}^T \mathbf{u_2}}{\|\mathbf{u_1}\| \|\mathbf{u_2}\| \tau}\right)$ denotes the function for computing the similarity between the vectors $\mathbf{u_1}$ and $\mathbf{u_2}$, and includes a temperature parameter τ. κ and σ denote the peak value and variance of the gaussian kernel.

We also present an ablation study on phase classification and overall action recognition in Supp. Sec. C.

E.2 Open-World Semi-Supervised Learning

Standard Semi-Supervised Framework: In the standard semi-supervised action recognition framework, the dataset consists of two sets:

- **Labeled Set** ($\mathbb{D}_l$): Includes video instances $\mathbf{v}^{(i)}$ and their corresponding action labels $\mathbf{y}^{(i)}$, from a set of predefined classes C.
  Formally, $\mathbb{D}_l = \{(\mathbf{v}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N_l}$.
- **Unlabeled Set** ($\mathbb{D}_u$): Contains unlabeled video instances that are assumed to belong to the same set of classes C.
  It is defined as $\mathbb{D}_u = \{\mathbf{v}^{(i)}\}_{i=1}^{N_u}$.

Open-World Extension: In the open-world semi-supervised learning framework, we introduce the presence of novel action classes within the unlabeled data:

- **Labeled Set**: Remains unchanged, with instances from the known classes C.
- **Unlabeled Set** ($\mathbb{D}'_u$): Now includes instances from both the known classes C and additional novel classes C_{novel}. Thus, samples in $\mathbb{D}'_u$ may belong to either C or C_{novel}. Represented as $\mathbb{D}'_u = \{\mathbf{v}^{(i)}\}_{i=1}^{N'_u}$.

The objective is to improve action recognition for classes in C using both $\mathbb{D}_l$ and $\mathbb{D}'_u$, while effectively handling the label noise from novel class instances C_{novel} in $\mathbb{D}'_u$.

Experimental Setup: For our experiments (Sec 4.4 of main paper) with the Diving48 dataset, 40 classes are designated as known classes C and the remaining 8 as novel classes C_{novel}. This setup tests the model's ability to not only accurately recognize actions from the known classes using the available data but also adapt to the presence of novel class instances.

F Detailed Comparison to Prior Work

F.1 Utilization of Alignment-Based Objective in Limited Labeled Setup

To the best of our knowledge, the work most closely related to ours in terms of utilizing an alignment cost is [2], which employs alignment cost directly to match queries with a support set in few-shot procedural video classification.

Our approach, however, differs from [2] significantly in several key aspects:

1. **Focus on Temporally Fine-Grained Actions:** We target temporally fine-grained actions where learning action phases is crucial, as opposed to procedural videos. Additionally, our semi-supervised framework leverages a substantial amount of unlabeled data, whereas [2] confines itself to a few-shot learning setup without using unlabeled data.
2. **Application of Alignability Score:** Instead of directly using the alignment cost for classification, we introduce a learnable alignability score to address a binary classification problem, encouraging a focus on intra-video features. Our concept of 'alignability' (determining if two clips are alignable) contrasts with the approach in [2], which applies alignment cost for multi-class classification.
3. **Temporal Context and Encoder Design:** [2] relies on a frame-level encoder and attempts frame-level alignment without temporal context. In contrast, our approach employs a frame-wise video encoder, pretrained with GITDL to grasp action phases before computing the alignment cost, thereby integrating temporal context into the model.
4. **Variant of DTW in [2]:** The study in [2] introduces an interesting variant of Dynamic Time Warping (DTW) with relaxed boundary conditions to find the optimal path of alignment. We explore this variant in our ablation study (Table 2). While it proves effective for procedural videos, in our context of temporally fine-grained actions, we observe that it performs less effectively than the regular DTW.

F.2 SSL Pretraining - GITDL

To utilize the unlabeled set $\mathbb{D}_u$ for learning a frame-wise video encoder f_A that focuses on intra-video dynamics such as action phases, we introduce the Gaussian Infused Temporal Distinctiveness Loss (GITDL). The most closely related SSL pretraining methods to our GITDL are [5] and [31].

Table 9: Different SSL Objectives for Alignability encoder

SSL pretraining of f_A	**10%**	**20%**
CARL	36.2	59.3
GITDL	**37.6**	**60.4**

Key differences include:

1. **Temporal Distinctiveness in GITDL vs. Temporal Invariance in CARL:** Our GITDL aims to learn explicit 'temporal distinctiveness', contrasting with the SSL objective of CARL, which promotes 'temporal invariance'. Mathematically, our loss (Eq. 1) considers only temporally-aligned frames as positives, whereas [5] treats all frames as positives (Eq. 1 of [5]), thereby fostering temporal invariance. Moreover, we apply a Gaussian prior to the negatives of the anchor, while CARL treats all negatives uniformly.

   Video as a Process in VSP: VSP([31]) views videos as a process and learns through a Brownian bridge with a triplet loss, which differs from our GITDL.
2. **Global and Local Clip Views:** Our approach incorporates both global and local views of a clip, providing more temporal context compared to the fixed-length clips in [5] and [31]. This global perspective better suits the subsequent learning stages, particularly the video-level alignability-verification objective using labeled data.

We have integrated the publicly available code of CARL ([5]) into our framework for comparative analysis, shown in Table 9. Although CARL yields impressive results on class-agnostic intra-video tasks (Table 7), it is slightly less effective in video-level semi-supervised tasks. Our conjecture is that this is due to the absence of global temporal context in [5] pretraining.

F.3 Training Cost Comparison with Prior Work

Our method only utilizes the single RGB modality, in contrast to methods like [26, 28], which employ additional modalities such as optical flow or temporal gradients. These extra modalities lead to a significant increase in training time due to two main factors: (1) the extended preprocessing time required to compute flow or temporal gradients, and (2) increased I/O overhead for loading both RGB and flow/gradient data. For example, computing optical flow for the Kinetics400 dataset can span several days and requires 3-5 terabytes of additional storage space. Conversely, our method efficiently operates on RGB-only videos, avoiding these extensive computational demands.

In terms of memory requirements, our framework is notably more efficient. Each branch of our model (f_E and f_A) is trained independently, thereby reducing the overall memory consumption. This is in stark contrast to training frameworks like [15,30], which necessitate running both teacher and student branches in training mode simultaneously, significantly increasing the memory footprint. Additionally, our approach does not require high-capacity teacher models. For instance, [30] employs a 3D-ResNet50-4x width as a teacher, whereas [9, 30] use two 3D-ResNet50 teachers. In comparison, our model efficiently utilizes only one teacher, further enhancing our method's resource efficiency.

References

1. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9922–9931 (2020)
2. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10618–10627 (2020)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Chang, C.Y., Huang, D.A., Sui, Y., Fei-Fei, L., Niebles, J.C.: D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3546–3555 (2019)
5. Chen, M., Wei, F., Li, C., Cai, D.: Frame-wise action representations for long videos via sequence contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13801–13810 (2022)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
7. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: International conference on machine learning. pp. 894–903. PMLR (2017)
8. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: Tclr: Temporal contrastive learning for video representation. Computer Vision and Image Understanding p. 103406 (2022). `https://doi.org/https://doi.org/10.1016/j.cviu.2022.103406`, `https://www.sciencedirect.com/science/article/pii/S1077314222000376`
9. Dave, I.R., Rizve, M.N., Chen, C., Shah, M.: 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 (2023)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1801–1810 (2019)
12. Goyal, R., Kahou, S.E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense (2017)
13. Hadji, I., Derpanis, K.G., Jepson, A.D.: Representation learning via global temporal alignment and cycle-consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11068–11077 (2021)
14. Haresh, S., Kumar, S., Coskun, H., Syed, S.N., Konin, A., Zia, Z., Tran, Q.H.: Learning by aligning videos in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5548–5558 (2021)
15. Jing, L., Parag, T., Wu, Z., Tian, Y., Wang, H.: Videossl: Semi-supervised learning for video classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1110–1119 (2021)

16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
17. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 513–528 (2018)
18. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. pp. 527–544. Springer (2016)
19. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3163–3172 (October 2021)
20. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11205–11214 (2021)
21. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 1134–1141. IEEE (2018)
22. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2616–2625 (2020)
23. Singh, A., Chakraborty, O., Varshney, A., Panda, R., Feris, R., Saenko, K., Das, A.: Semi-supervised action recognition with temporal contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10399 (2021)
24. Thoker, F.M., Doughty, H., Bagad, P., Snoek, C.G.: How severe is benchmark-sensitivity in video self-supervised learning? In: European Conference on Computer Vision. pp. 632–652. Springer (2022)
25. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
26. Xiao, J., Jing, L., Zhang, L., He, J., She, Q., Zhou, Z., Yuille, A., Li, Y.: Learning from temporal gradient for semi-supervised action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3252–3262 (2022)
27. Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., Jiang, Y.G.: Svformer: Semi-supervised video transformer for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18816–18826 (2023)
28. Xiong, B., Fan, H., Grauman, K., Feichtenhofer, C.: Multiview pseudo-labeling for semi-supervised learning from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7209–7219 (2021)
29. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2949–2958 (2022)
30. Xu, Y., Wei, F., Sun, X., Yang, C., Shen, Y., Dai, B., Zhou, B., Lin, S.: Cross-model pseudo-labeling for semi-supervised action recognition. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2959–2968 (2022)
31. Zhang, H., Liu, D., Zheng, Q., Su, B.: Modeling video as stochastic processes for fine-grained video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2225–2234 (2023)
32. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the IEEE international conference on computer vision. pp. 2248–2255 (2013)
33. Zou, Y., Choi, J., Wang, Q., Huang, J.B.: Learning representational invariances for data-efficient action recognition. arXiv preprint arXiv:2103.16565 (2021)