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/

Abstract. Real-life applications of action recognition often require a fine-grained understanding of subtle movements, e.g., in sports analytics, user interactions in AR/VR, and surgical videos. Although fine-grained actions are more costly to annotate, existing semi-supervised action recognition has mainly focused on coarse-grained action recognition. Since fine-grained actions are more challenging due to the absence of scene bias, classifying these actions requires an understanding of action-phases. Hence, existing coarse-grained semi-supervised methods do not work effectively. In this work, we for the first time thoroughly investigate semi-supervised fine-grained action recognition (FGAR). We observe that alignment distances like dynamic time warping (DTW) provide a suitable action-phase-aware measure for comparing fine-grained actions, a concept previously unexploited in FGAR. However, since regular DTW distance is pairwise and assumes strict alignment between pairs, it is not directly suitable for classifying fine-grained actions. To utilize such alignment distances in a limited-label setting, we propose an Alignability-Verification-based Metric learning technique to effectively discriminate between fine-grained action pairs. Our learnable alignability score provides a better phase-aware measure, which we use to refine the pseudo-labels of the primary video encoder. Our collaborative pseudo-labeling-based framework '*FinePseudo*' significantly outperforms prior methods on four fine-grained action recognition datasets: Diving48, FineGym99, FineGym288, and FineDiving, and shows improvement on existing coarse-grained datasets: Kinetics400 and Something-SomethingV2. We also demonstrate the robustness of our collaborative pseudo-labeling in handling novel unlabeled classes in open-world semi-supervised setups.

Keywords: Fine-grained Action Recognition · Semi-supervised learning

1 Introduction

Considering the action recognition problem in practice, many critical applications demand high precision in classifying subtle movements. For instance, in analyzing surgical videos to monitor subtle patient movements [52], AR and VR

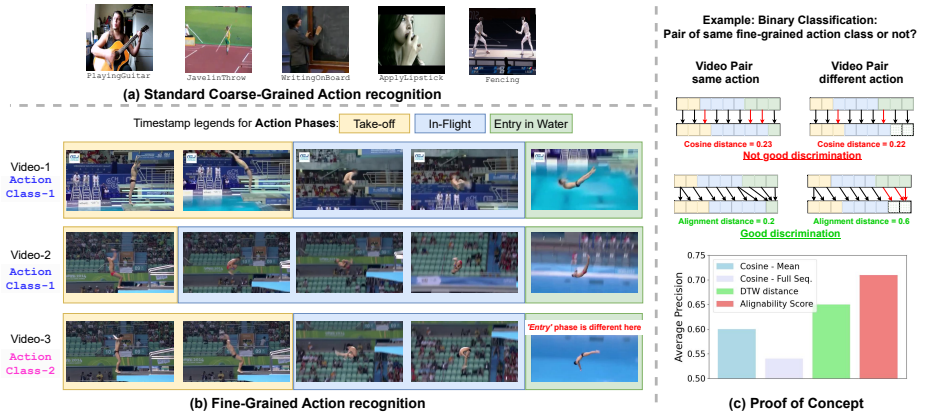

Fig. 1: **(a)** Sample actions from standard coarse-grained action recognition dataset (UCF101) **(b)** Sample actions from fine-grained action recognition dataset (Diving48) **(c)** For proof-of-concept, we choose a binary classification problem of fine-grained actions, where the model has to predict whether the pair of videos belong to the same class or not. We consider Diving48 dataset with 10% training data. We first obtain the frame-wise video embedding from a pretrained framewise video-encoder model (Details in Sec. 3.3). The top part of (c) shows that the cosine distance computed at each timestamp does not provide a discriminative measure, whereas, DTW-based alignment cost provides a clear difference in pair of same vs different classes. The bottom part of (c), shows the performance of the binary classification task in terms of average precision, where our alignability-score significantly outperforms the other standard distances.

applications [58], require identifying the user's nuanced movements for a more responsive interaction, and in sports analytics [27,34], it enables detailed action quality assessment and injury prevention.

Although fine-grained action recognition (FGAR) allows for wider adoption of action recognition in real-life applications, research has mainly focused on coarse-grained action recognition [19,22,28,46,66]. For instance, from Fig. 1(a), we observe that coarse-grained action covers broader classes, such as 'PlayingGuitar' vs 'JavelinThrow'. Subtle human movements are not essential for classifying these, given their very different motion pattern and inherent scene bias (i.e., the scene provides substantial cues for identifying action) [13]. Conversely, Fig. 1(b) illustrates fine-grained action categories from 'Diving', comprising *action-phases* like 'Take-off', 'In-flight', and 'Entry into Water'. This figure demonstrates that even a difference in the 'Entry' phase from video-2 to video-3 alters the action class from class-1 to class-2. This suggests that FGAR can significantly benefit from learning action-phases.

However, annotating such fine-grained actions poses significant challenges. Unlike coarse-grained actions, fine-grained actions require extensive, often repetitive viewing and expert annotation, making the process time-consuming and costly. This underscores the need for a semi-supervised learning approach for FGAR. However, current semi-supervised methods, designed for broader ac-

tion categories, heavily rely on complex augmentation schemes like strongly and weakly augmented versions [59], token-mix [55], or actor-cutmix augmentations [68]. These techniques, while successful in standard datasets mainly for exploiting scene bias, may not be effective for FGAR due to scene uniformity across actions. Moreover, recent video-level self-supervised methods [18], successful in limited data contexts, do not effectively capture frame-level changes in action phases, which is crucial for recognizing fine-grained actions.

To build a solution tailored for fine-grained action recognition, we conduct a preliminary study to better understand the efficacy of different distance metrics in differentiating fine-grained videos. Let's take the example of binary classification of fine-grained actions, shown in Fig. 1(c). Here, the goal is to verify whether the two videos belong to the same or different class utilizing the embeddings from a frame-wise video encoder (f_A) in a limited labeled data setting. Our experiments demonstrate that standard feature distances like cosine distance are inadequate for this classification task. Particularly, we notice that computing cosine-distance over the temporally pooled features loses the temporal-granularity whereas computing cosine distance on the temporally unpooled features is suboptimal since different action phases take different amounts of time, e.g., phases in video-1 and video-2 of Fig. 1(b). Therefore, a better way to compute the distance between a pair of fine-grained actions should be done by making *phase-to-phase* comparisons. One way to obtain such phase-aware distance is by aligning the phases of the video embeddings. Hence, we hypothesize that alignability (i.e., whether two videos are alignable or not) based verification can provide a better phase-aware solution to differentiate fine-grained actions.

One way to achieve such phase-aware distance is through alignment distance - dynamic time warping (DTW) optimal path distance. We see a significant boost in class-discrimination capability over regular cosine distances as shown in Fig.1(c) bar chart. This observation has not been explored before to solve FGAR in the limited labeled setting. At the same time, standard DTW distance is not an ideal class-discriminative measure as its optimal path assumes strict alignment between two videos and the final distance depends on the length of the video and frame-level similarities. Based on this observation, we propose an *alignability-verification*-based metric learning technique to learn from the labeled data and produce a *learnable alignability score* for a pair of videos. In the chart Fig.1(c), we see that our learnable alignability score improves the class-discriminative capability of DTW and provides a better distance measure for discriminating a pair of fine-grained videos.

Once such limited-labeled training is completed, we can utilize this alignability metric for pseudo-labeling (PL) by producing class-wise alignability-scores. These temporal aliganbility based pseudo-labels provide complementary information to the standard pseudo-labels generated from output confidence scores. To benefit from these complementary sets of pseudo-labels, we employ a collaborative pseudo-labeling process for semi-supervised fine-grained action recognition. Particularly, we combine the class predictions from frame-wise encoder, f_A,

and finetuned video encoder, f_E, to get a refined pseudo-label. We update these pseudo-labels iteratively and conduct training in a self-training framework.
The major contributions of this work are summarized as follows:

- Our work is the first to thoroughly study the problem of fine-grained semi-supervised action recognition. We present *FinePseudo*, a co-training framework where we utilize temporal-alignability to improve the pseudo-labeling process of unlabeled fine-grained videos.
- To learn effectively from the limited labeled fine-grained videos, we propose a alignability-verification-based metric learning technique.
- For collaborative pseudo-labeling, we design a non-parametric classifier-based prediction from the learnable alignability scores to refine output predictions.
- We evaluate our method on 4 fine-grained action recognition datasets: Diving48, FineGym99, FineGym288, and FineDiving, where our method significantly outperforms prior semi-supervised action recognition methods. Our method also performs competitively against the prior methods on coarse-grained datasets like Kinetics400 and Something-SomethingV2.
- We demonstrate the robustness of our collaborative pseudo-labelling method in handling novel unlabeled classes in open-world semi-supervised setups.

2 Prior Work

Semi-supervised Action Recognition Semi-supervised learning is still a growing area of research for action recognition compared to the image domain [1,2,6,12,39,42,43,53,61,64,67]. To exploit the additional temporal dimension, various methods have employed additional modalities, including temporal gradients [54], optical-flow [56], and P-frames [50]. Concurrently, interesting augmentation schemes have been proposed, such as slow-fast streams [51], strong-weak augmentations [59], CutMix [68], and token-mix [55]. While [18] shows the potential of self-supervised video representations (videoSSL) in leveraging the unlabeled videos for semi-supervised setup.

However, these approaches mainly address semi-supervised action recognition problems focusing on coarse-grained actions with significant scene bias [13], where the scene context provides substantial cues for action recognition. For fine-grained actions, which typically occur within the same scene, methods tailored for scene-bias datasets may not be fully applicable. For instance, augmentations like token-mix or CutMix might lose their effectiveness in uniform scene environments. Similarly, some methods may be partially ineffective, such as the appearance branch of [56], or the temporally-invariant teacher of [18]. While approaches like [47] have shown results on [24], their application has not been thoroughly explored beyond human-object interaction, leaving a gap in addressing diverse human actions.

Motivated by these gaps, we propose, for the first time, a dedicated semi-supervised action recognition framework that not only achieves state-of-the-art performance for fine-grained action classes but also performs comparably or better in standard coarse-grained action recognition. Categorically, our method

is a pseudo-label-based technique building upon existing videoSSL representations. Our method introduces a novel approach for pseudo-label generation using temporal-alignability-verification-based decisions, which provides a fresh perspective in the semi-supervised action recognition domain. Additionally, our method demonstrates increased robustness to open-world problems, a dimension not previously explored in semi-supervised action recognition. This robustness further distinguishes our approach from the constrained focus of prior work.

Fine-grained Video understanding There is another line of work that specifically focuses on intra-video dynamics for learning class-agnostic downstream tasks like action-phase classification, Kendall's tau [21], Aligned Phase Agreement [16] etc. Some recent works have demonstrated the learning of powerful fine-grained intra-video representations in a weakly-supervised manner [3, 21, 25, 26] and even on unlabeled data utilizing intra-video self-supervised techniques like [10, 17, 32, 44, 65].

Interestingly, some of these works use alignment-based training objectives to resolve class-agnostic tasks [9, 25, 26, 60]. However, they strictly assume that videos are 'alignable' (from the same action class) and do not explore leveraging the alignment property across video classes to learn data-efficient fine-grained action classification. In contrast to the typical 'alignment' objective, we opt for an 'alignability' objective, where we decide if an unlabeled video belongs to a fine-grained class based on how well it aligns with the limited labeled samples. To the best of our knowledge, we are the first to leverage 'alignability'-based intra-video representations in the video-level action recognition task in a semi-supervised setup. For a detailed comparison with prior work, refer Supp. Sec.F.

3 Method

In semi-supervised action recognition, a limited labeled set $\mathbb{D}_l = \{(\mathbf{v}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N_l}$ comprising video instances and their associated action labels is employed alongside a significantly larger unlabeled set $\mathbb{D}_u = \{\mathbf{v}^{(i)}\}_{i=1}^{N_u}$. The goal is to leverage both these sets to enhance the performance of action recognition.

Our framework, FinePseudo, is a pseudo-labeling-based co-training approach, as depicted in the schematic diagram in Fig. 3. It mainly consists of two branches: (1) Action encoder f_E responsible for learning high-level video-semantics features such as actions and (2) Auxiliary alignability-encoder f_A which is a frame-wise video encoder - video transformer network (VTN) [35], to focus on learning the low-level intra-video representations stemming from action phases.

In this section, we provide more method details of our framework. First, f_A is trained through alignability-verification-based metric learning from the labeled data (Sec.3.1). Then, for pseudo-labeling from the unlabeled data, the trained f_A is utilized to provide learned alignability scores for each class, which are passed through a non-parametric classifier to obtain classwise predictions. This alignability-based prediction from f_A is combined with the prediction from the regular video encoder f_E to obtain a collaborative pseudo-label, which is used

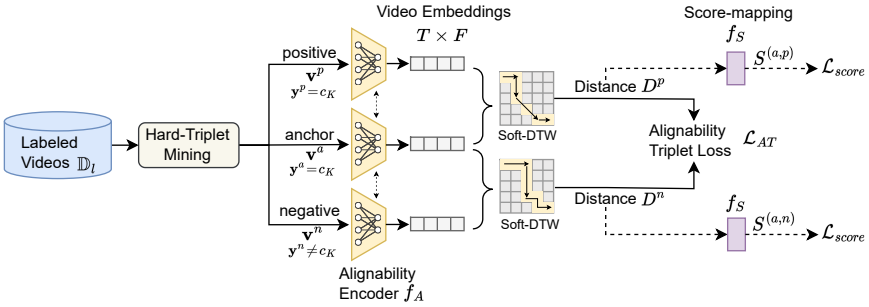

Fig. 2: Alignability-Verification based Metric Learning is proposed to is proposed to decide how well two video instances are alignable and produce an 'alignability score' for effective learning from a limited labeled set $\mathbb{D}_l$. Our approach employs a triplet loss ($\mathcal{L}_{AT}$), considering videos from identical action classes as positive and those from different classes as negative. We selectively mine hard-negatives from the sampled minibatch based on alignment distance, presenting a challenging learning task for the model f_A. Additionally, we incorporate a matching loss $\mathcal{L}_{score}$ to quantify the alignment between videos, serving as a verification task to determine whether a video pair belongs to the same class (i.e. alignable or target label = 1) or different classes (i.e. non-alignable or target label = 0). Further details are provided in Sec. 3.1.

for the self-training process (Sec.3.2). A complete algorithm for our FinePseudo training is provided in Sec. 3.3.

3.1 Alignability-Verification based Metric Learning

The underlying hypothesis is that video instances from the same action class are more alignable compared to those from different classes (as seen in Fig. 1(c)). The objective of this training stage is to solve the alignability verification task, which determines how well two videos are alignable. This knowledge is critical for producing a reliable 'learnable alignability score' for a pair of labeled and unlabeled video instances, subsequently aiding in the improvement of pseudo-label quality through a non-parametric classifier within the self-training paradigm.

In our approach, class labels are utilized in learning the alignability-verification task which is a binary classification problem, distinct from the regular multi-class classification setting [7]. This strategy encourages the network to focus on differentiating pairs from the same or different classes based on their alignment distance, promoting the learning of more class-agnostic intra-video features.

Alignment Cost: Consider a pair of videos U and V, each with T frames. To compute the alignment cost between these videos, they are processed through the f_A network, yielding frame-wise video embeddings $\mathbf{u}$ and $\mathbf{v}$, each of shape $T \times F$, where F represents the output feature size of f_A. An element-wise cost matrix $\mathbb{C} \in \mathbb{R}^{T \times T}$ is constructed, with each element computed using the cosine distance: $\mathbb{C}(i,j) = h(\mathbf{u}(i), \mathbf{v}(j))$. To identify the optimal alignment path, softDTW [14], a differentiable variant of the dynamic time warping algorithm [4], is utilized.

The softDTW distance, $D(\mathbf{u}, \mathbf{v})$, is then calculated using the following recursive formula:

$$D(\mathbf{u}, \mathbf{v}) = \mathbb{C}(i, j) + \gamma\text{-smooth-min}(\Pi_{\text{cost}}(i, j)) \tag{1}$$

The function γ-smooth-min performs a differentiable minimum operation of the possible costs $\Pi_{\text{cost}}(i, j)$ from the point (i, j) along the paths $(i, j-1)$, $(i-1, j)$, and $(i-1, j-1)$. Now, we utilize this alignment-cost as the distance to build our metric learning training objective.

Alignability-verification Triplet loss: For a labeled instance $\mathbf{V}^{(i)}$ of class $y^{(i)} = c_K$, we select another instance $\mathbf{V}^{(j)}$ of the same class as positive and an instance $\mathbf{V}^{(k)}$ from a different class as negative. After obtaining the video-embeddings, the alignability-based triplet loss is computed as follows:

$$\mathcal{L}_{AT} = \sum_{i=1}^{N} \Big[D(\mathbf{v}^{(i)}, \mathbf{v}^{(j)}) - D(\mathbf{v}^{(i)}, \mathbf{v}^{(k)}) + m \Big] \tag{2}$$

where D denotes the softDTW distance, m is the margin of the triplet loss, and N is the number of samples in the mini-batch B. Hard-negative mining is employed from the same mini-batch B for constructing these triplets, with further analysis in Supp. Sec. C.

Learnable Alignability Score: Finally, to determine the alignability of video pairs based on their alignment cost, we propose a normalized scale ranging from 0 (not alignable) to 1 (fully alignable). The computed distance D is mapped through a non-linear scaling function f_S and passed through a sigmoid activation (ς) to yield a learnable Alignability-score S between any sequence embeddings $\mathbf{u}$ and $\mathbf{v}$.

$$S(\mathbf{u}, \mathbf{v}) = \varsigma(f_S(D(\mathbf{u}, \mathbf{v}))) \tag{3}$$

To train this scaling function, a binary cross-entropy loss function is employed:

$$\mathcal{L}_{Score} = -[y_A \log(S(\mathbf{u}, \mathbf{v})) + (1 - y_A) \log(1 - S(\mathbf{u}, \mathbf{v}))] \tag{4}$$

Where y_A label is assigned 0 for the negative pair and 1 for the positive pair. The overall training objective for our alignability-verification-based metric learning can be expressed as:

$$\mathcal{L}_{AV} = \mathcal{L}_{AT} + \omega \mathcal{L}_{Score} \tag{5}$$

where, ω hyperparameter is the relative weighting factor.

While the alignability encoder (f_A) is trained through the alignability-verification training from the labeled set $\mathbb{D}_l$, the action encoder (f_E) is trained through regular cross-entropy loss as shown in the equation below:

$$\mathcal{L}_{CE}^{(i)} = -\sum_{c=1}^{N_c} \mathbf{y}_c^{(i)} \log \mathbf{p}_c^{(i)} \tag{6}$$

Where N_c is the number of classes, $y_c^{(i)}$ is the ground-truth class and $\mathbf{p}_c$ is the classwise prediction by classification head of f_E.

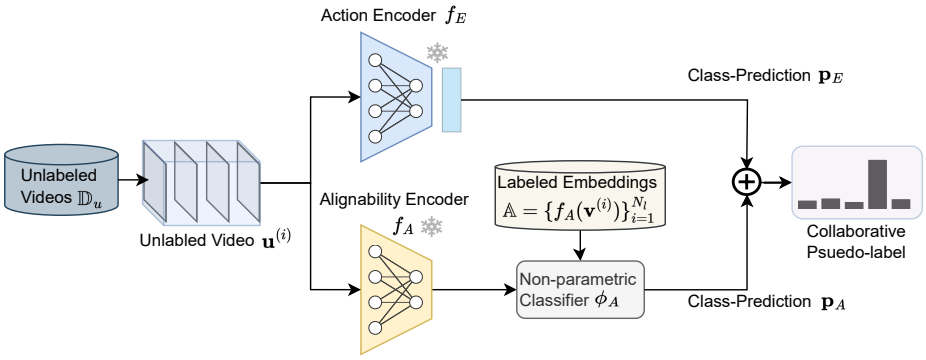

Fig. 3: Collaborative Pseudo-labeling: The unlabeled instance $\mathbf{u}^{(i)}$ undergoes processing by both video encoders (f_E and f_A). For the Action Encoder f_E, its prediction ($\mathbf{p}_E$) is derived via its classification head. For the Alignability Encoder f_A, the embedding of $\mathbf{u}^{(i)}$ computes class-wise alignability scores against a gallery of labeled embeddings $\mathbb{A}$. These scores are then used to generate a class-wise prediction $\mathbf{p}_A$ using the non-parametric classifier ϕ_A. As these predictions stem from distinct supervisory signals—$\mathbf{p}_E$ from video-level and $\mathbf{p}_A$ from alignability-based supervision—they offer complementary insights, resulting in a refined collaborative pseudo-label.

3.2 Collaborative Pseudo-Labeling

Once both action encoder f_E and alignability encoder f_A is trained with the $\mathbb{D}_l$, they are utilized to generate pseudo-labels for the videos of unlabeled set $\mathbb{D}_u$. Before we start pseudo-labeling, we first construct a set $\mathbb{A}$ by obtaining embedding of video of D_l by passing it through encoder f_A. This process is formalized as $\mathbb{A} = \{f_A(\mathbf{v}^{(i)})\}_{i=1}^{N_l}$

For an unlabeled video $U \in \mathbb{D}_u$, its embedding $\mathbf{u}$ is obtained by passing it through the alignability encoder f_A. The alignability score for each class c in the labeled dataset is computed by randomly sampling ρ embeddings from $\mathbb{A}$ corresponding to class c, denoted as $\mathbb{A}_c^{\rho}$. The average alignability score $\bar{S}_c$ for class c is calculated as:

$$\bar{S}_c = \frac{1}{\rho} \sum_{\mathbf{a} \in \mathbb{A}_c^{\rho}} S(\mathbf{u}, \mathbf{a}) \tag{7}$$

For computing the class prediction $\mathbf{p}_A$ for the unlabeled video U using the softmax function with a temperature parameter τ. This function is applied to the alignability scores, yielding the class prediction as:

$$\mathbf{p}_A(c) = \frac{\exp(\bar{S}_c/\tau)}{\sum_j \exp(\bar{S}_j/\tau)} \tag{8}$$

The denominator in Eq. 8 sums over all classes j in $\mathbb{D}_l$, producing a probability distribution over the classes and indicating the predicted likelihood of the unlabeled video U belonging to each class. Since there is no parameter involved

in getting the prediction $\mathbf{p}_A$ we can call it non-parametric classifier ϕ_A of the alignability encoder.

The same unlabeled video U is passed through f_E and its classifier head to obtain its class prediction $\mathbf{p}_E$. The overall final prediction $\mathbf{p}$ for a video U is obtained by adding the predictions from both classifiers: $\mathbf{p} = \mathbf{p}_A + \mathbf{p}_E$. We apply a confidence threshold θ to each prediction $\mathbf{p}$. If the highest confidence score in the prediction $\mathbf{p}$ exceeds the threshold θ, the sample is considered for generating a hard pseudo-label; otherwise, the sample is discarded. In this way, we achieve refined pseudo-labels and they are used for the next iteration of labeled training for both f_A and f_E.

3.3 Algorithm

Let's consider the action encoder model f_E and the alignability model f_A, parameterized by θ_E and θ_A, respectively. In our semi-supervised training framework, firstly we employ our novel GITDL-based self-supervised pretraining (Details in Supp. Sec. E) on the unlabeled dataset $\mathbb{D}_u$ to learn frame-wise video representations focusing on intra-video dynamics, and secondly, leveraging both labeled $\mathbb{D}_l$ and pseudo-labeled data in a collaborative self-training process. These steps are put together in Algorithm 1, which outlines the complete process for our *FinePseudo* framework for semi-supervised action recognition.

4 Experiments

4.1 Datasets and Metrics

Diving48 [31] is a fine-grained dataset on competitive diving, with 48 distinct patterns across roughly 18k videos. Each class underscores the intricacies of a diver's movements, stressing the need for detailed temporal analysis to capture subtle differences in takeoff, flight, and entry phases.
FineGym [45] is a large-scale, fine-grained action recognition dataset that provides hierarchical annotations for four different gymnastic events: Vault, Floor Exercise, and Balance Beam. It comprises two main splits: FineGym99 with 99 actions from 29k videos, and FineGym288 with 288 actions from 32k videos.
FineDiving [57] dataset comprises diverse diving events, covering 52 action classes across 23 difficulty degrees.
Kinetics400 [8] encompasses 400 human action classes across approximately 260k videos sourced from YouTube.
Something-SomethingV2 [24] is another large dataset with clips that are object class agnostic, focusing on a wide range of 174 hand-object interactions.
For further dataset details, refer Supp. Sec. A.
Evaluation Metric: Following standard protocols in prior work [18, 59], we evaluate 3 independent label splits and report the mean Top-1 accuracy.
For implementation details, refer Supp. Sec. B

Algorithm 1: FinePseudo Training Algorithm

Inputs:
  Datasets: $\mathbb{D}_u$, $\mathbb{D}_l$
  #Epochs: max_epoch_ssl, $max\_epoch\_labeled$, max_iter, max_epoch_st
  Learning Rates: α_A, α_E
  Hyperparameters: Confidence threshold θ
Output: Action Encoder model θ_E

SSL Pretraining on Unlabeled Set $\mathbb{D}_u$:
for $e_0 \leftarrow 1$ **to** max_epoch_ssl **do**
  $\theta_A \leftarrow \theta_A - \alpha_A \nabla \mathcal{L}_{GITDL}(\theta_A)$ (Refer Supp. Eq. 1)
end

Training from the Labeled Set $\mathbb{D}_l$:
for $e_0 \leftarrow 1$ **to** $max\_epoch\_labeled$ **do**
  $\theta_E \leftarrow \theta_E - \alpha_E \nabla \mathcal{L}_{CE}(\theta_E)$ (Refer Eq. 6)
end
for $e_0 \leftarrow 1$ **to** $max\_epoch\_labeled$ **do**
  $\theta_A \leftarrow \theta_A - \alpha_A \nabla \mathcal{L}_{AV}(\theta_A)$ (Refer Eq. 5)
end

Self-Training through Collaborative Pseudo-Labeling:
for $iter \leftarrow 1$ **to** max_iter **do**
  for *each sample in* $\mathbb{D}_u$ **do**
    Obtain combined class-prediction $\mathbf{p} = \text{avg}(\mathbf{p}_A, \mathbf{p}_E)$
      Predicted class $\hat{\mathbf{y}}$
    if *confidence of* $\hat{\mathbf{y}} > \theta$ **then**
      Add (sample, predicted label $\hat{\mathbf{y}}$) to $\mathbb{D}_l$
    end
  end
  for $epoch_0 \leftarrow 1$ **to** max_epoch_st **do**
    $\theta_E \leftarrow \theta_E - \alpha_E \nabla \mathcal{L}_{CE}(\theta_E)$
    $\theta_A \leftarrow \theta_A - \alpha_A \nabla \mathcal{L}_{AV}(\theta_A)$
  end
end

4.2 Evaluation on Fine-grained datasets

In order to maintain comparability across methods, we utilize the R2plus1D-18 network. We compare various baselines such as video self-supervised methods [11, 15, 36–38], classical semi-supervised learning baselines [29, 41], and state-of-the-art video semi-supervised methods [18, 54, 68] in Table 1. In the first section of Table 1, we study video self-supervised methods by taking their publicly available Kinetics400 self-supervised weights and fine-tuning them for the fine-grained action recognition task under limited labeled data. We observe that the methods [15] and [11], which explicitly promote temporal distinctiveness, perform better than other video self-supervised methods.

Based on this observation, we use the best-performing SSL weights [15] for **all semi-supervised methods** in the second part of Table 1. Firstly, we note

Table 1: Comparison with state-of-the-art semi-supervised methods on Fine-grained Action recognition datasets under various % of labeled data setting. Highlighted **Red** shows the best results and Blue shows second best results. All results are reported on R2plus1D-18 utilizing the exact same amount of training data.

Method	Diving48			FineGym99			FineGym288			FineDiving		
	5%	10%	20%	5%	10%	20%	5%	10%	20%	5%	10%	20%
TCLR CVIU'22 [15]	14.3	33.1	53.7	43.2	64.2	74.9	36.0	56.8	67.2	23.2	42.3	65.2
VidMoCo CVPR'21 [37]	12.6	31.4	52.5	41.6	62.8	73.8	34.2	55.8	66.8	21.9	40.6	64.8
GDT ICCV'21 [38]	12.2	31.7	51.8	42.0	62.0	73.3	35.3	56.0	66.6	21.2	40.9	64.3
AVID CVPR'21 [33]	10.0	30.4	51.5	40.3	60.3	72.7	32.5	55.6	64.5	20.6	39.6	62.7
RSPNet AAAI'21 [11]	14.0	33.0	53.7	43.4	64.0	75.2	36.8	56.4	67.1	23.0	42.5	65.1
PL ICML'13 [30]	14.4	33.4	54.0	43.2	64.4	75.1	34.9	55.5	67.1	23.5	42.0	66.1
UPS ICLR'21 [41]	14.6	33.6	54.1	-	-	-	-	-	-	-	-	-
ActorCM CVIU'22 [68]	14.7	33.8	54.7	43.8	65.0	75.9	36.5	56.9	67.7	-	-	-
TG-FM CVPR'21 [54]	16.0	33.8	54.4	44.1	64.9	75.7	36.9	56.6	67.6	-	-	-
TimeBal CVPR'23 [18]	15.8	33.7	56.3	44.4	65.9	76.1	37.3	57.8	68.6	25.1	43.9	67.5
Ours (*FinePseudo*)	**20.9**	**37.6**	**60.4**	**49.2**	**69.9**	**80.0**	**41.7**	**62.5**	**73.4**	**28.4**	**46.8**	**71.9**

that classical semi-supervised baselines, namely PL and UPS, do not perform as well compared to video semi-supervised methods. Our method consistently outperforms all prior methods by an absolute **4-5%** in terms of top-1 accuracy.

Evaluating with Transformer architecture Since the AIM-ViTB architecture [62] achieves state-of-the-art performance on the Diving48 dataset in a fully-supervised setting, we find it interesting to base our comparisons. In this architecture, the ViT-B backbone [20] is kept frozen and initialized with the CLIP [40] visual encoder, and spatio-temporal adaptor layers are trained.

Firstly, using this architecture (Table 2), we observe a significant improvement compared to Table 1. Next, examining the results of recent semi-supervised methods [18, 55], it becomes evident that token-mix augmentations from [55] are not as effective in fine-grained datasets as in coarse-grained ones. Similarly, videoSSL-based semi-supervised methods like [18] also underperform due to the ineffectiveness of some components like temporally-invariant teacher in fine-grained datasets. Our method achieves a clear improvement of **4-5%**, *demonstrating its potential to further enhance the strong foundational model pretraining for fine-grained action recognition in a limited labeled setting.*

Table 2: Results with AIM model on Diving48 dataset

Method	5%	10%	20%
Supervised	37.28	55.33	75.36
PL [29]	37.33	55.40	75.42
UPS [41]	37.70	55.61	75.56
SVFormer [55]	38.00	56.02	76.20
TimeBal [18]	38.12	55.80	76.01
Ours	**43.02**	**60.79**	**80.02**

Table 3: Results on standard Coarse-grained Action recognition datasets at various % of labeled set. Highlighted **Red** shows the best and Blue shows second best results.

Method	Backbone	Params (M)	ImgNet init?	#F	Kinetics400			S. SomethingV2		
					1%	5%	10%	1%	5%	10%
MT NeuRIPS'17 [49]	TSM-ResNet18	13	✗	8	6.8	23.0	-	7.3	20.2	30.2
S4L ICCV'19 [63]	TSM-ResNet18	13	✗	8	6.3	23.3	-	7.2	18.6	26.0
MM NeuRIPS'19 [5]	TSM-ResNet18	13	✗	8	7.0	21.9	-	7.5	18.6	25.8
FM NeuRIPS'20 [48]	TSM-ResNet18	13	✗	8	6.4	25.7	-	6.0	21.7	33.4
TCL CVPR'21 [47]	TSM-ResNet18	13	✗	8	11.6	31.9	-	9.9	31.0	41.6
TG-FM CVPR'21 [54]	3D-ResNet18	13.5	✗	8	9.8	-	43.8	-	-	-
MvPL ICCV'21 [56]	3D-ResNet18	13.5	✗	8	5.0	-	36.9	-	-	-
CMPL CVPR'22 [59]	3D-ResNet18	13.5	✗	8	16.5	-	53.7	-	-	-
TimeBal CVPR'23 [18]	3D-ResNet18	13.5	✗	8	17.1	-	54.9	-	-	-
Ours (*FinePseudo*)	3D-ResNet18	13.5	✗	8	**18.6**	**43.2**	**56.1**	**13.1**	**34.3**	**45.4**
FM NeuRIPS'20 [48]	SlowFast-R50	60	✗	8	10.1	-	49.4	6.5	25.3	37.4
MvPL ICCV'21 [56]	3D-ResNet50	31.8	✗	8	17.0	-	58.2	-	-	-
CMPL CVPR'22 [59]	3D-ResNet50	31.8	✗	8	17.6	-	58.4	-	-	-
TimeBal CVPR'23 [18]	3D-ResNet50	31.8	✗	8	19.6	-	61.2	-	-	-
SVFormer CVPR'23 [55]	T.Former(ViT-S)	30.7	✗	16	17.2	42.3	58.1	9.9	31.7	42.9
Ours (*FinePseudo*)	3D-ResNet50	31.8	✗	8	**21.4**	**47.5**	**62.6**	**13.4**	**34.7**	**46.1**

4.3 Evaluation on Coarse-grained action datasets

Although the focus of our work is on the evaluation of fine-grained actions, we also evaluate coarse-grained action datasets as shown in Table 3. For comparability, results are presented using two backbones: 3D-ResNet18 and 3D-ResNet50 [23], with an input resolution of 224×224 and 8-frame clips. Our learnable-alignability score-based approach shows favorable or slightly improved performance over prior best methods across both backbones. This demonstrates that our approach, not reliant on a strict alignment criterion, generalizes well for generic coarse-grained human actions and is not confined to fine-grained actions.

4.4 Evaluation on Open-World setting

In previous evaluations, it was assumed that the unlabeled data belonged to one of the classes in the labeled set. However, in practical scenarios, an unlabeled sample could originate from any *novel* (unknown) action class. Refer to Supp. Sec. E for more details about this protocol.

Table 4: Results with open-world setting on Diving48 dataset. All models are R2plus1D-18.

Method	10%	20%
Supervised	39.60	50.23
Pseudo-labeling	38.29	49.41
UPS [41]	38.93	49.56
TimeBalance [18]	39.90	50.88
Ours	**42.21**	**55.37**

To explore the open-world setting, we utilize the Diving48 dataset, where 40 classes are randomly selected as *known* classes and the remaining 8 classes are designated as *novel* classes. For this protocol, we consider the R2plus1D-18 model with SSL initialization from Kinetics400 from [15], and

Table 5: Ablation with different components of framework

	Action Encoder	Alignability Encoder f_A			Top-1 Accuracy	
		SSL ($\mathbb{D}_U$)	Metric Learning ($\mathbb{D}_L$)			
	f_E	$\mathcal{L}_{GITDL}$	$\mathcal{L}_{AT}$	$\mathcal{L}_{Score}$	10%	20%
(PL)	✓	-	-	-	33.40	54.00
(a)	✓	-	-	-	33.10	53.70
(b)	✗	✓	✓	✓	32.82	51.05
(c)	✓	✓	✗	✗	33.50	53.76
(d)	✓	✓	✓	✗	33.73	55.67
(e)	✓	✓	✗	✓	36.11	59.32
(f)	✓	✗	✓	✓	35.23	58.64
(g)	✓	✓	✓	✓	**37.64**	**60.40**

the results are reported in Table 4. The supervised baseline, which only utilizes the labeled data from the 40 classes, is established for comparison. The regular pseudo-label setting degrades the performance of the supervised baseline, as the novel unlabeled samples introduce noise during self-training. The prior best semi-supervised method [18] also fails to show noticeable improvement over the supervised baseline, as its teacher model categorizes the unlabeled sample into one of the known classes before distillation to a student. In contrast, our approach, with its non-parametric classification in PL generation, effectively filters out unknown classes based on low alignability scores, thereby achieving improvement over other methods. For additional results, refer Supp. Sec. D.

4.5 Ablation Study

We demonstrate the ablation experiments on Diving48 dataset with R2plus1D-18 network by default. Additional ablations and detail in Supp. Sec. C.

Evaluating Contributions of Training Components: In Table 5, we study the effect of each training step in our framework: SSL pretraining on $\mathbb{D}_u$ and Alignability-based metric learning on $\mathbb{D}_l$.

- When using individual video encoders (`Rows a, b`), f_E performs better than f_A, however, it is significantly suboptimal compared to their collaborative use in (`g`). `Row (PL)` shows regular PL baseline [29] for the f_E which helps only by a small margin. The Alignability-Verification-based metric learning significantly help to improve the capability of recognizing fine-grained actions.(`Row c vs Row g`).
- `Row d,e vs Row g` suggest that both alignability-triplet loss and score loss contribute towards the final performance. Since $\mathcal{L}_{AT}$ provides a more challenging task with hard triplets and margin, it helps significantly compared to the simpler binary classification objective of $\mathcal{L}_{Score}$.
- Proposed GITDL self-supervised pretraining for f_A helps 2% on the final performance (`Row f vs Row g`).

Pseudo-Label Refinement Strategies: We examine the impact of various pseudo-labeling (PL) strategies on the Diving48 dataset with a limited labeled split, as shown in Table 6. Alongside the final performance, we also report the number of pseudo-labels (PLs) that surpass the threshold and their accuracy, as determined by comparison with the ground truth in the fully labeled set.

In the first section, we explore standard pseudo-labeling methods based on the model's class prediction confidence and uncertainty. We find that incorporating uncertainty with confidence (as shown in the second row) enhances PL accuracy but reduces the quantity of PLs. Because of this reduction, the improved PL accuracy does not translate into a noticeable gain in final performance.

Table 6: Psuedo-Label refinement methods.

Pseudo-Labelling(PL)	**PL statistics**		**Results**	
Method	**Count**	**Acc.**	**10%**	**20%**
Regular- Conf. based	4813	85.4	33.40	53.95
Uncertainty based	3565	87.9	33.58	54.07
Label verification	1981	97.0	37.09	59.57
Non-Parametric Classif.	4558	96.4	**37.64**	**60.40**

In the third row, we introduce an alignability-score based PL verification strategy. After a class prediction by f_E clears the confidence-based threshold, we calculate the alignability score for its predicted class. If this score exceeds an alignability-score threshold (set at 0.6), we accept the PL for self-training; otherwise, it is discarded. This alignability-based score verification significantly improves PL accuracy and consequently enhances overall performance.

Finally, in the last row, we present results using our combined class prediction approach, which incorporates a prediction $\mathbf{p}_A$ obtained through a non-parametric (NP) classifier (as detailed in Sec. 3.2). This method substantially increases the count of PLs over the verification-based PL approach and improves the overall results.

5 Conclusion and Future Work

We present FinePseudo, a novel co-training-based semi-supervised framework tailored for fine-grained action recognition. Our framework effectively utilizes the strengths of a coarse-level video encoder dedicated to high-level action understanding, alongside a frame-wise video encoder focusing at capturing low-level intra-video dynamics, particularly action phases. Notably, FinePseudo improves existing state-of-the-art video SSL methods and foundational models when trained for semi-supervised learning for fine-grained action recognition. The efficacy of our collaborative pseudo-labeling process is further validated in open-world semi-supervised scenarios.

For future work, exploring multi-modal temporal-alignability, such as video and audio integration, could enhance the efficiency of semi-supervised action recognition. Additionally, the potential of FinePseudo extends to other video understanding tasks requiring fine-grained temporal understanding, like action quality assessment *etc*.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) and Center for Smart Streetscapes (CS3) under NSF Cooperative Agreement No. EEC-2133516.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
2. Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., Rabbat, M.: Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8443–8452 (2021)
3. Bansal, S., Arora, C., Jawahar, C.: My view is the best view: Procedure learning from egocentric videos. In: European Conference on Computer Vision. pp. 657–675. Springer (2022)
4. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd international conference on knowledge discovery and data mining. pp. 359–370 (1994)
5. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems 32, pp. 5049–5059. Curran Associates, Inc. (2019)
6. Cai, Z., Ravichandran, A., Favaro, P., Wang, M., Modolo, D., Bhotika, R., Tu, Z., Soatto, S.: Semi-supervised vision transformers at scale. Advances in Neural Information Processing Systems **35**, 25697–25710 (2022)
7. 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)
8. 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)
9. 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)
10. 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)
11. Chen, P., Huang, D., He, D., Long, X., Zeng, R., Wen, S., Tan, M., Gan, C.: Rspnet: Relative speed perception for unsupervised video representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1045–1053 (2021)
12. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in Neural Information Processing Systems **33** (2020)
13. Choi, J., Gao, C., Messou, J.C., Huang, J.B.: Why can't i dance in the mall? learning to mitigate scene bias in action recognition. Advances in Neural Information Processing Systems **32** (2019)

14. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: International conference on machine learning. pp. 894–903. PMLR (2017)
15. 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
16. Dave, I.R., Caba, F., Shah, M., Jenni, S.: Sync from the sea: Retrieving alignable videos from large-scale datasets. In: European Conference on Computer Vision (2024)
17. Dave, I.R., Jenni, S., Shah, M.: No more shortcuts: Realizing the potential of temporal self-supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1481–1491 (2024)
18. 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)
19. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L.: Large scale holistic video understanding. In: European Conference on Computer Vision. pp. 593–610. Springer (2020)
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
21. 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)
22. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
23. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 6202–6211 (2019)
24. 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)
25. 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)
26. 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)
27. Hong, J., Fisher, M., Gharbi, M., Fatahalian, K.: Video pose distillation for few-shot, fine-grained sports action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9254–9263 (2021)
28. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV) (2011)
29. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks (2013)

30. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
31. 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)
32. 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)
33. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021)
34. Naik, B.T., Hashmi, M.F., Bokde, N.D.: A comprehensive review of computer vision in sports: Open issues, future trends and research directions. Applied Sciences **12**(9), 4429 (2022)
35. 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)
36. Newell, A., Deng, J.: How useful is self-supervised pretraining for visual tasks? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7345–7354 (2020)
37. 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)
38. Patrick, M., Asano, Y.M., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations (2021)
39. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11557–11568 (2021)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
41. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: International Conference on Learning Representations (2021)
42. Rizve, M.N., Kardan, N., Khan, S., Shahbaz Khan, F., Shah, M.: Openldn: Learning to discover novel classes for open-world semi-supervised learning. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI. pp. 382–401. Springer (2022)
43. Rizve, M.N., Kardan, N., Shah, M.: Towards realistic semi-supervised learning. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI. pp. 437–455. Springer (2022)
44. 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)

45. 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)
46. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. CoRR **abs/1804.09626** (2018), http://arxiv.org/abs/1804.09626
47. 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)
48. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
49. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
50. Terao, H., Noguchi, W., Iizuka, H., Yamamoto, M.: Compressed video ensemble based pseudo-labeling for semi-supervised action recognition. Machine Learning with Applications p. 100336 (2022)
51. Tong, A., Tang, C., Wang, W.: Semi-supervised action recognition from temporal augmentation using curriculum learning. IEEE Transactions on Circuits and Systems for Video Technology (2022)
52. Tscholl, D.W., Rössler, J., Said, S., Kaserer, A., Spahn, D.R., Nöthiger, C.B.: Situation awareness-oriented patient monitoring with visual patient technology: A qualitative review of the primary research. Sensors **20**(7), 2112 (2020)
53. Wang, J., Lukasiewicz, T., Massiceti, D., Hu, X., Pavlovic, V., Neophytou, A.: Npmatch: When neural processes meet semi-supervised learning. In: International Conference on Machine Learning. pp. 22919–22934. PMLR (2022)
54. 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)
55. 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)
56. 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)
57. 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)
58. Xu, X., Mangina, E., Campbell, A.G.: Hmd-based virtual and augmented reality in medical education: a systematic review. Frontiers in Virtual Reality **2**, 692103 (2021)
59. 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)

60. Xue, Z., Grauman, K.: Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
61. Yang, F., Wu, K., Zhang, S., Jiang, G., Liu, Y., Zheng, F., Zhang, W., Wang, C., Zeng, L.: Class-aware contrastive semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14421–14430 (2022)
62. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: Aim: Adapting image models for efficient video understanding. In: International Conference on Learning Representations (2023)
63. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1476–1485 (2019)
64. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems **34**, 18408–18419 (2021)
65. 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)
66. Zhao, H., Torralba, A., Torresani, L., Yan, Z.: Hacs: Human action clips and segments dataset for recognition and temporal localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8668–8678 (2019)
67. Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C.: Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14471–14481 (2022)
68. Zou, Y., Choi, J., Wang, Q., Huang, J.B.: Learning representational invariances for data-efficient action recognition. arXiv preprint arXiv:2103.16565 (2021)