How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs

Anonymous CVPR submission

Paper ID 63

Abstract

We aim to understand how actions are performed and identify subtle differences, such as 'fold firmly' vs. 'fold gently'. To this end, we propose a method which recognizes adverbs across different actions. However, such fine-grained annotations are difficult to obtain and their long-tailed nature makes it challenging to recognize adverbs in rare action-adverb compositions. Our approach therefore uses semi-supervised learning with multiple adverb pseudo-labels to leverage videos with only action labels. Combined with adaptive thresholding of these pseudo-adverbs we are able to make efficient use of the available data while tackling the long-tailed distribution. Additionally, we gather adverb annotations for three existing video retrieval datasets, which allows us to introduce the new tasks of recognizing adverbs in unseen action-adverb compositions and unseen domains. Experiments demonstrate the effectiveness of our method, which outperforms prior work in recognizing adverbs and semi-supervised works adapted for adverb recognition. We also show how adverbs can relate fine-grained actions. This paper has been accepted to CVPR 2022.

1. Introduction

This paper aims to recognize fine-grained differences between actions such as whether a person is swimming *slowly* or *quickly* or cutting *evenly* or *unevenly*. Understanding how actions are performed is key to understanding the actions themselves and their outcomes. Improved perception of the action manner would allow both humans and robots to better imitate actions, as well as better discrimination between fine-grained action categories, where the difference can simply be how much an object moves [16]. Previous works can address the question of *what* is happening in a video [76], *when* an action is happening [64], *who* is performing an action [69] and *where* it is taking place [39]. However, very few works have looked at *how* actions happen, as we do in this paper.

In language, how an action is performed can be described with adverbs, thus we focus on recognizing such adverbs. Two works have previously investigated adverb recognition [10, 47]. However, these works either focus on adverbs describing facial expressions and moods [47] or only studied a handful of adverbs [10] limiting the ways to answer "how is the action being performed?". This highlights a key challenge in learning adverbs and more generally fine-grained video understanding: the time-consuming data collection. The more subtle the differences between videos, the more difficult it is to collect a large amount of labels. To address these challenges and better describe how actions are performed, we scale up the number of adverbs which can be learned by utilizing videos with only action labels. Furthermore, multiple adverbs can co-occur and apply to the same action. We can thus better learn adverbs from videos in a semi-supervised fashion by obtaining extra adverb labels through multi-adverb pseudo-labeling (see Fig. 1).

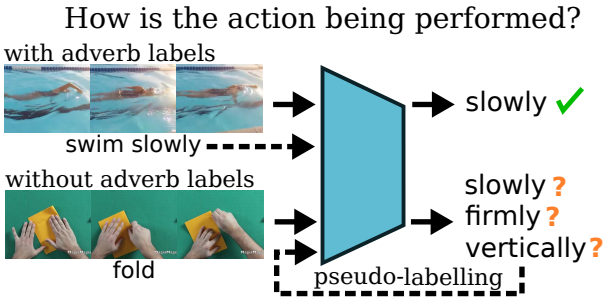

Figure 1. We answer how actions are happening by learning adverbs of different actions. We do this in a semi-supervised manner where we use action-only videos with multi-adverb pseudo-labeling.

As our main contribution, we propose to reformulate the adverb recognition problem as a semi-supervised learning problem. In Sec. 3, we propose the first method for semi-supervised learning of adverbs, in which we apply multiple adverb pseudo-labels to actions and use an adaptive threshold to cope with the long-tail distribution of adverbs. In Sec. 4, we create several new adverb recognition benchmarks by automatically mining action-adverb pairs from the captions in existing video retrieval datasets [25, 61, 66]. Alongside this we propose two new tasks for addressing how actions happen: first recognizing adverbs in unseen compositions and second in recognizing adverbs across domains. In Sec. 5, we demonstrate our multi-adverb pseudo-labeling approach obtains a considerable improvement over prior works in recognizing seen compositions of verbs and adverbs as well as improving generalization in these new tasks.

2. Related Work

We first review works focused on fine-grained understanding of actions followed by video retrieval. We then examine works which have focused specifically on adverbs. Finally, we look at semi-supervision for other vision tasks.

Fine-grained Action Understanding. Recent datasets have focus on fine-grained actions [8,16,32,55]. For instance, in FineGym [55] a model must distinguish between 'salto forwards' and 'salto backwards'. While some actions are similar, the majority of works [6, 12, 26, 33, 36, 59, 60, 62, 74] model actions as distinct categories leaving the model to implicitly learn similarities. Some works instead explicitly model actions as compositions of components, either through sub-actions [48,49] or verbs and noun combinations with [22, 41] or without [3,8,40,56,77] the spatial location of the noun. We instead, look at fine-grained differences between actions by recognizing adverbs in combination with different verbs.

Other works recognize actions through combinations of specified attributes [35, 52, 53, 70, 72]. For instance, with Temporal Query Networks, Zhang *et al.* [72] propose to determine the correct attributes by first attending to the most relevant video parts with an attribute-focused query. The attributes studied in these work do not consider adverbs, instead they indicate the presence of an object, a person's pose or the number of repetitions of an action.

Video Retrieval. Potentially more fine-grained than action recognition is video-text retrieval, which aims to retrieve the correct caption describing the video. The majority of such works create sentence-level features with recurrent networks [9, 19, 44], learned pooling [42] or transformers [13, 37, 68, 75]. While retrieval datasets [19, 25, 29, 46, 61, 66] do contain adverbs, models use verbs and nouns to distinguish videos as they are more frequent [63]. Rather than relying on a sentence encoding to indicate the distinctive elements of a caption, some prior works focus on certain parts of speech [7, 63, 67]. Again, the focus is on verbs and nouns, with Wray *et al.* [63] learning separate embeddings for each and Chen *et al.* [7] learning a hierarchical text encoding from verbs, nouns and the semantic relation between them. We instead focus on understanding adverbs and how these apply to different verbs. We obtain new, more varied, action-adverb annotations from three video retrieval datasets.

Adverbs. Some works have studied individual adverbs. For instance, Benaim *et al.* [4] identify whether videos are played *quickly*, Epstein *et al.* [11] recognize whether an event occurred *accidentally* and Heidarivincheh *et al.* [18] pinpoint when an action has finished *completely*. There are two prior works which look at recognizing adverbs more generally. Pang *et al.* [47] propose a fully-supervised method using video, pose and expression features. The adverbs in this work focus primarily on moods and expressions such as *solemnly* and *excitedly*. Doughty *et al.* [10] learn adverbs from weak supervision with attention locating the video segments relevant to the action. Adverbs are then learned as transformations in an embedding. This approach is still label-hungry, requiring sufficient adverb-labeled actions for all action-adverb compositions. We instead utilize action-only labeled videos to recognize adverbs in both seen and unseen compositions. For this we introduce three new adverb datasets, significantly increasing the number of adverbs from 6 to 34 and the number of compositions from 263 to 1,550.

Semi-supervision. Many strategies have been explored for semi-supervised learning such as pseudo-labeling [1, 17, 27], consistency regularization [2, 5, 57, 58], generative models [45, 50] and fine-tuning self-supervised models [71]. For instance, Lee [27] propose an efficient method for pseudo-labeling where one-hot labels are obtained for an unlabeled sample by taking the highest confidence prediction. Sohn *et al.* propose the consistency regularization approach FixMatch [58], where the loss aims to make the label predicted for two augmented versions of an image consistent.

Several works focus on semi-supervised learning for video [15,23,57,65]. TCL by Singh *et al.* [57] maximizes the prediction similarities between different speeds of a video. Xiong *et al.* [65] target consistency in the pseudo-labels predicted by RGB, optical flow and temporal gradient streams. Gavrilyuk *et al.* [15] also propagate pseudo-labels between modalities, but instead aim to distill motion information so downstream tasks only need the RGB modality in training.

Since these works target image, object or action recognition, they are unsuitable for adverbs. Adverbs are compositional both with actions and other adverbs and these compositions have a long-tailed distribution. We propose a semi-supervised approach to learn adverbs via multi-adverb pseudo-labeling and adaptive thresholding to address these challenges. We also demonstrate how our approach can improve generalization to unseen action-adverb compositions.

3. Semi-supervised Learning of Adverbs

Our work aims to understand how an action is being performed in a video by predicting the adverb(s) applicable to that action. An overview of our approach can be seen in Fig. 2. Labeled data can be used to learn to recognize adverbs in composition with different actions (Sec. 3.1). However, a key challenge in understanding subtle differences, such as adverbs, is lack of labeled data. In this work we propose to better learn adverbs with semi-supervised learning by creating pseudo-adverb labels for video clips with action labels only (Sec. 3.2). We observe that multiple adverbs can apply to the same action, therefore we propose to better utilize the available data with our multi-adverb pseudo-labeling (Sec. 3.3). Another challenge is the long-tailed nature of adverbs. We use adaptive thresholding on the adverb pseudo-labels so that our approach is effective on all adverbs, not only the most frequent (Sec. 3.4).

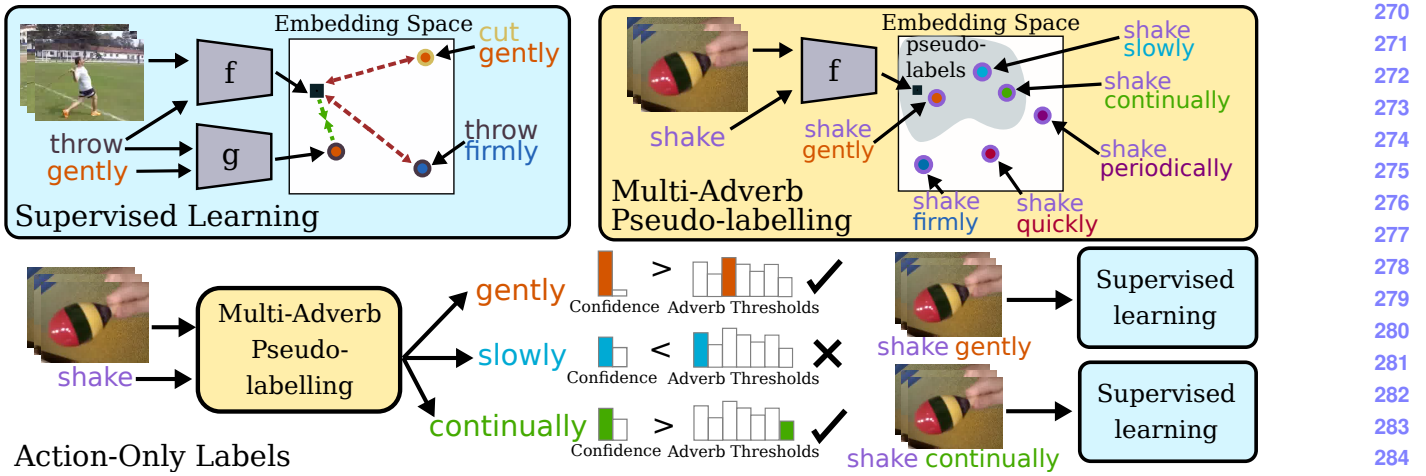

Figure 2. Semi-supervised learning of adverbs. In the supervised case our input is the video with an action and adverb label, *e.g.* throw *gently*. f embeds the video parts relevant to the action close to the ground-truth action-adverb text embedding from g. For videos without adverb labels we create multiple pseudo-labels by finding the most confident adverb predictions when compared to their antonym. In this example, for the action shake we obtain the pseudo-adverbs *slowly*, *gently* and *continually*. We use per-adverb thresholds to select which of these pseudo-labels we should use. Each video is then trained with the selected pseudo-labeled adverbs as if they were in the supervised set.

3.1. Learning Adverbs with Labeled Data

Given a video clip $x \in X$ and a label for the action of interest $a \in A$, the goal of adverb recognition is to correctly predict the adverb $\hat{m}$ which applies to action a. Since many adverbs are not mutually exclusive and can simultaneously apply to the same action, we particularly focus distinguishing between the labeled adverb m and its antonym $\text{ant}(m)$. In the supervised case we learn to recognize adverbs with a labeled set of videos $S{=}\{(x, a, m)\}$.

As in prior work [10], we learn adverbs in a video-text embedding space as this allows actions and adverbs to be compositional. The goal is to embed the parts of the video relevant to the action close to a text representation of that action modified by the adverb. Specifically, we learn a visual embedder $f : X, A \rightarrow E$ and a textual embedder $g : A, M \rightarrow E$. We aim for $f(x, a)$ and $g(a, m)$ to be close in the embedding space E and $f(x, a)$ to be far from embeddings with other actions $g(a', m)$ and with the antonym adverb $g(a, \text{ant}(m))$. We do this with two triplet losses:

$$\mathcal{L}_{act}(S) = \sum_{(x,a,m)\in S} \max(0, \text{dist}(f(x,a), g(a,m)) - \text{dist}(f(x,a), g(a',m)) + \gamma_1) \quad s.t.\ a \neq a', \tag{1}$$

$$\mathcal{L}_{adv}(S) = \sum_{(x,a,m)\in S} \max(0, \text{dist}(f(x,a), g(a,m)) - \text{dist}(f(x,a), g(a,\text{ant}(m))) + \gamma_2), \tag{2}$$

where dist is a distance metric and γ_1 and γ_2 are margins.

3.2. Pseudo labeling Adverbs

Now we consider how we can improve the learning of adverbs by utilizing video clips with only action labels. To make use of this type of data we propose to pseudo label action clips with adverbs. Formally, we have a video set without adverb labels $U{=}\{(x, a)\}$. For each video clip x with action label a, we can create a single adverb pseudo-label $\widetilde{m}$ by selecting the adverb in the closest text representation as the pseudo-label. Let $d(x, a, m){=}\text{dist}(g(a, m), f(x, a))$:

$$\widetilde{m} = m_n \ \text{ where } n = \text{argmin}_i\ d(x, a, m_i), \tag{3}$$

where m_n is a label indicating adverb n. For the action-only videos we can then use $\widetilde{m}$ in place of m in L_{adv} (Equation 2). This gives us the overall loss:

$$\mathcal{L} = \mathcal{L}_{act}(S) + \mathcal{L}_{adv}(S) + \mathcal{L}_{act}(U) + \mathcal{L}_{adv}(U). \tag{4}$$

3.3. Multi-Adverb Pseudo-Labeling

While actions in the supervised set S are labeled with a single adverb, the majority of adverbs are not mutually exclusive, meaning multiple adverbs can apply to a single action. We thus propose multi-adverb pseudo labeling. To do this we take the top k most confident adverbs and let the adverb pseudo-label $\widetilde{m}$ to be a set of pseudo-labels:

$$\widetilde{m} = \{m_n\} \text{ s.t. } n \in \text{topk}_i(\text{conf}(x, a, m_i)), \tag{5}$$

where

$$\text{conf}(x, a, m) = \frac{e^{d(x,a,m)}}{e^{d(x,a,m)} + e^{d(x,a,\text{ant}(m))}}. \tag{6}$$

With this definition of $\text{conf}(x, a, m)$ we take the most confident to mean the greatest relative difference between the adverb and its antonym rather than the closest adverbs.

Now we have multiple adverb pseudo labels for each of the action-only labeled videos in U. We optimize for each pseudo-labeled adverb, meaning the overall loss becomes:

$$\mathcal{L} = \mathcal{L}_{act}(S) + \mathcal{L}_{adv}(S) + \mathcal{L}_{act}(U) + \sum_{\widetilde{m}} \mathcal{L}_{adv}(U). \tag{7}$$

3.4. Adaptive Adverb Thresholding

The problem of recognizing adverbs is naturally long-tailed. Not only are some adverbs much more common than others, but certain compositions of actions and adverb are also more frequent. Using our multi-adverb pseudo-labeling we are able to make better use of the available data. However, it has a tendency to only select the most frequent adverbs, as the adverbs it is most confident in selecting are in the action-adverb pairs with the most examples.

We take inspiration from semi-supervised object detection where the long-tail is also present [31] and propose to use adaptive thresholding. The threshold is dynamically adjusted for each adverb m. Not only does this mean that the threshold is increased for the more confident adverbs so that fewer noisy pseudo-labels are used, but importantly the threshold is lowered for the adverbs with fewer confident predictions, meaning they are no longer unrepresented in the pseudo-labels. We adapt an initial threshold τ to an adverb-specific threshold τ_m as follows:

$$\tau_m = \left(\frac{\sum_{U:m\in\widetilde{m}} \text{conf}(x,a,m)}{\frac{1}{N}\sum_U |\widetilde{m}|} \right)^{\lambda} \tau, \qquad (8)$$

where N is the number of adverbs. The sum of confidence scores for an adverb m, $\sum_{U:m\in\widetilde{m}} \text{conf}(x,a,m)$, acts as an approximation of the model's overall confidence for predicting this adverb over its antonym. We then divide this by the average number of pseudo-labels per adverb. λ is a smoothing factor which controls the amount the model focuses on underrepresented adverbs. With λ=0, all adverbs use the original threshold τ. The adverb-specific threshold τ_m is applied to filter the available pseudo-labels, so that only the pseudo labels with $\text{conf}(x,a,m) > \tau_m$ for $m \in \widetilde{m}$ are used.

4. Adverb Datasets and Tasks

We evaluate our approach on **HowTo100M Adverbs** [10] which mined adverbs from 83 tasks in HowTo100M [43]. Since the annotations were obtained from the automatically transcribed narrations of instructional videos, they are noisy; in training ∼44% of the annotated action-adverb pairs are not visible in the video clip. The dataset contains 5,824 clips annotated with action-adverb pairs from 72 verbs and 6 adverbs. A clear limitation of this dataset is the small number of adverbs it contains, we thus create three new adverb datasets from existing video retrieval datasets: VATEX Adverbs, MSR-VTT Adverbs and ActivityNet Adverbs. These contain less noise and a greater variety of adverbs.

4.1. Adverb Annotations from Video Captions

We extract verb-adverb annotations for videos in existing video-text datasets to obtain three new adverb datasets. From available datasets [14, 19–21, 25, 28–30, 34, 46, 51, 54, 61, 66, 73] we find VATEX [61], ActivityNet Captions [25] and MSR-VTT [66] contain the best variety of adverbs with sufficient instances. Each contains video clips with corresponding text captions. VATEX consists of 35k 10 second video clips, each of which has 10 English captions, resulting in a total of 260k captions. In MSR-VTT each clip is 10-30 seconds and has 20 captions giving a total of 10k clips and 200k captions. ActivityNet Captions contains 20k videos with an average of 3.65 temporally localized sentences per video, resulting in a total of 100k clips and matching captions. Each dataset was sourced from YouTube, thus some videos are no longer available. At the time of collection we obtained: 32,161 video clips for VATEX, 7,511 for ActivityNet and 5,197 for MSR-VTT.

Figure 3. Example video clips and action-adverb annotations.

Extracting Adverb Annotations. To extract adverb annotations from the captions in these datasets we search for adverbs and their corresponding verbs. We use SpaCy's English core web model where RoBERTa [38] performs Part-of-Speech tagging and dependency parsing on each caption. We search for verbs which have adverbs as children, excluding any verbs with a negative dependency to another word. We filter out non-visual adverbs, adverbs whose antonym doesn't appear and adverbs which appear less than 10 times or only appear in combination with a single action. The resulting verbs and adverbs from the three datasets are manually clustered, starting with the clusters from [10]. This process forms 137 verb clusters and 34 adverb clusters in 17 adverb-antonym pairs. Fig. 3 shows examples of the video clips alongside the discovered action-adverb pairs.

Adverb Datasets. This results in three adverb datasets: **VATEX Adverbs**, **ActivityNet Adverbs** and **MSR-VTT Adverbs**. Table 1 shows statistics of each. VATEX Adverbs is the largest with 34 adverbs appearing across 135 actions to form 1,550 unique action-adverb pairs. The distribution of actions, adverbs and their compositions are heavily long-tailed (see Fig. 4). Each dataset considers many more adverbs than the existing HowTo100M Adverbs which contains only 6. We measure the quality of each dataset's annotations with a 200 video sample. Since the new datasets come from human written captions, where a person has explicitly chosen the adverb to describe the action, the annotations are much less noisy than HowTo100M Adverbs.

Dataset	Adverbs & Actions				Videos		Tasks		
	Adverbs	Actions	Pairs	Accuracy	Clips	Length (s)	Seen	Unseen	Domain
HowTo100M Adverbs [10]	6	72	263	44.0%	5,824	20.0	✓	-	-
VATEX Adverbs	34	135	1,550	93.5%	14,617	10.0	✓	✓	Source
MSR-VTT Adverbs	18	106	464	91.0%	1,824	15.7	✓	-	Target
ActivityNet Adverbs	20	114	643	89.0%	3,099	37.3	✓	-	Target

Table 1. Our three newly proposed adverb datasets have more adverbs, actions, unique pairs and higher annotation accuracy than HowTo100M Adverbs [10] and also allow us to study recognition of adverbs in unseen action-adverb compositions and new domains.

4.2. Adverb Recognition Tasks

With these datasets we aim to learn to recognize each adverb in combination with different actions. As in prior work [10], we want to recognize adverbs in previously seen action-adverb compositions. We additionally propose two new adverb recognition tasks: first in unseen compositions and second in unseen domains. We explain each below.

Task I: Seen Compositions. Adverbs and actions are compositional, an adverb $m \in M$ can apply to many different actions $a \in A$. Assume we have a set of action-adverb compositions $(a, m) \in C$. When recognizing adverbs in seen compositions, all compositions in the test set have been seen in the labeled training set, *i.e.* $C_{test} \subseteq C_{labeled}$. This tests whether the model can successfully compose and recognize adverbs across various actions. For this evaluation we use our newly proposed VATEX Adverbs as well as HowTo100M Adverbs [10]. To partition VATEX Adverbs into train and test we follow the original train and test split. This gives us 11,782 video clips in training and 2,835 in testing over the 34 adverbs. HowTo100M-Adverbs contains 6 adverbs and consists of 5,475 video clips in training and 349 in testing.

Task II: Unseen Compositions. To fully capture the compositional nature of actions and adverbs it is necessary for a model to generalize beyond seen compositions. We thus propose to recognize adverbs in unseen compositions, *i.e.* $C_{test} \cap C_{labeled} = \emptyset$. We focus on VATEX Adverbs for this since it has the most action-adverb pairs. We partition the pairs into two disjoint sets. For each action, both the pair with an adverb and its antonym are in the same set. Each set contains 50% of the pairs and every action and adverb is present in both sets. We take one split for training and further partition the second split, using half the instances of each pair as the test set and half as the action-only set.

Task III: Unseen Domains. Since a key challenge of fine-grained video understanding is the collection of labeled data, it is unreasonable to assume we will have labels in every domain where we wish to recognize adverbs. We therefore propose to test the transferability of learned adverbs to new domains. Here our labeled data S comes from a domain D_S while our test set and action-only labeled data U come from a distinct domain $D_U \neq D_S$. We use VATEX Adverbs as the source and MSR-VTT Adverbs and ActivityNet Adverbs as targets. We partition both targets in two 50% splits, one for testing and the other as the action-only labeled training data.

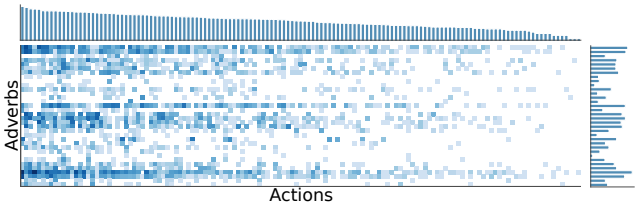

Figure 4. Distribution of action-adverb pairs in VATEX Adverbs shown on a log-scale. The adverbs and actions labels are long tailed as are their compositions. Labeled version in supplementary.

5. Experiments

We first describe the implementation details of our method and the evaluation metric used. We then analyze the contribution of our model's components and compare to semi-supervised baselines for recognizing adverbs in seen compositions. Finally, we evaluate our approach for recognizing adverbs in unseen compositions and unseen domains.

Implementation Details. All videos are sampled at 25fps and scaled to 256px in height. Each video is divided into 1-second segments with one 16-frame snippet extracted per segment. We use a frozen I3D network as the backbone, one for RGB and one for optical flow. The output of the global pooling layer for each modality is concatenated to create a $T \times 2048D$ feature, where T is the length of the video clip in seconds. The video embedder f uses transformer-style attention to locate the relevant video parts with the T video features as the keys and the action as the query. The text embedder g uses GloVe embeddings to represent the actions and learns adverbs as linear transformations on action embeddings. See [10] for more details. Optimization is done with Adam [24]. Models are trained with a supervised batch size of 128 and learning rate of 10^{-4} for 1000 epochs. As in [10] we introduce the adverbs after the 200th epoch, until that moment we train g as an action embedder. In experiments without thresholding, we reduce noise by letting the adverb representations train for 100 epochs before introducing pseudo-labels. The ratio of adverb-labeled to action-only labeled samples in a batch is the same as the total ratio. Unless otherwise specified, we set the maximum pseudo-labels

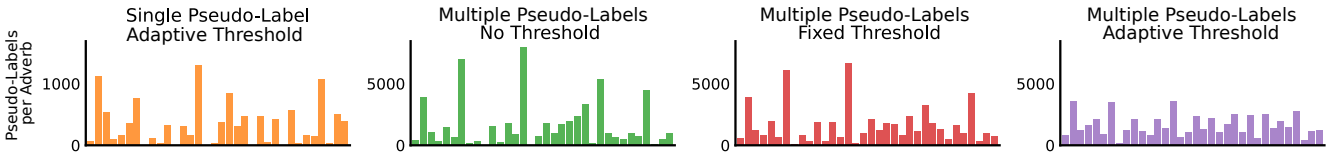

Figure 5. Distribution of adverb pseudo labels over all videos. Each bar indicates the number of videos pseudo-labeled with a particular adverb. With multi-adverb pseudo-labeling and adaptive thresholding in our model (purple), pseudo-labels are better distributed among the possible adverbs. With either single-adverb pseudo-labels (yellow) or other types of thresholding (green and red) the pseudo-labels reflect the long-tail distribution of ground-truth labels.

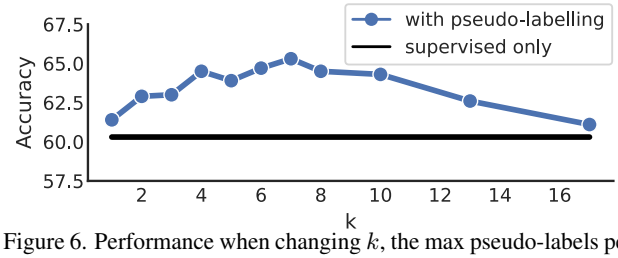

Figure 6. Performance when changing k, the max pseudo-labels per video. Using multi-adverb pseudo-labeling improves performance.

Method	Acc.
Closest	61.7
Max Difference	63.9

(a)

Thresholding	Acc.
None	61.1
Fixed	61.4
Adaptive	63.9

(b)

Table 2. (a) Pseudo-label selection. Considering antonyms with max difference is better than using the closest adverbs. (b) Type of thresholding. Adaptive thresholding gives better pseudo-labels.

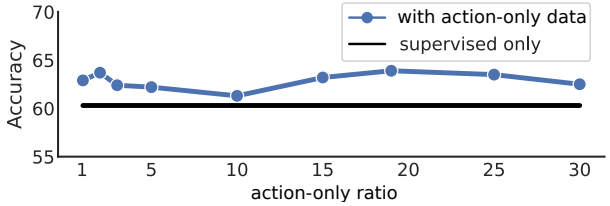

Figure 7. Effect of the ratio of adverb-labeled to action-only data. Any ratio of action-only data improves over using just labeled data.

per video to k=5 (Eq. 5), the base threshold to τ=0.6 (Eq. 8) and the smoothing factor to λ=0.1 (Eq. 8).

Evaluation Metric. We use adverb-antonym binary classification accuracy from [10]. That is the accuracy when considering the ground-truth adverb vs. its antonym. This suits the available ground-truth labels since they indicate a single adverb, while multiple adverbs may apply to an action. As the distributions of adverbs in our new datasets are long-tailed, we report the average accuracy over adverbs for these, rather than average over videos.

5.1. Ablation Study

We first perform several ablation studies evaluating the effect of each of the proposed model components. For these we recognize adverbs in seen compositions with VATEX Adverbs since this has the greatest variety of adverbs. Experiments are performed with 5% of the training set as the labeled set and the remainder as the action-only labeled set.

Multi-Adverb Pseudo-Labeling. Fig. 6 shows the effect of k, the maximum pseudo-labeled adverbs per video. We see using multiple pseudo-labels (k>1) offers a large advantage over supervised-only learning and semi-supervised learning with a single pseudo-label (k=1). The best performance is with k=7, although any value in the range $4 \geq k \geq 10$ is good. When allowing many different adverbs to apply to an action ($k \geq 13$) the performance drops, since this many adverbs rarely co-occur, although this is still better than supervised only learning.

Using multi-adverb pseudo-labeling allows us to make more efficient use of the data at our disposal as each video clip is used to learn multiple adverbs. It also encourages exploration of adverbs less frequent in the labeled set, which we show in Fig. 5. With a single adverb pseudo-label (yellow), the overall distribution of pseudo-labels is highly imbalanced and mimics the long-tail distribution of ground-truth labels. There are even 5 adverbs which have not been used in the pseudo-labels. Our multi-adverb pseudo-labeling (purple) reduces this bias, with pseudo-labels better distributed across the possible adverbs.

Pseudo-Label Selection. A standard approach to pseudo-labeling in an embedding space would be to take the closest embeddings as the pseudo-label(s). We instead take the adverbs with the greatest difference between the embedded video's proximity to the adverb modified action and the antonym modified action. We compare these approaches in Table 2a, which shows our approach improves the result.

Adaptive Thresholding. Table 2b compares the adaptive thresholding we use to no thresholding and a fixed threshold for all adverbs. The adaptive thresholding improves the result by 2.5% over fixed thresholding, which has little impact itself. With fixed thresholding, once the most common adverbs have exceeded this threshold the model will pseudo-label all actions with these adverbs, ignoring rarer adverbs. The adaptive thresholding allows the pseudo-label selection to be more balanced (as shown in Fig. 5).

Ratio of Action-Only Data We test the effect of the ratio of adverb-labeled to action-only labeled videos in Fig. 7. This shows training with any amount of action-only data gives better performance. We observe two peaks in Fig. 7. With an action-only ratio $\geq$15 the model is able to see all available

Method	VATEX Adverbs						HowTo100M Adverbs					
	1%	2%	5%	10%	20%	Av.	1%	2%	5%	10%	20%	Av.
Supervised only	54.0	54.5	60.3	64.7	64.2	59.5	67.3	**68.5**	67.9	73.4	74.8	70.4
Pseudo-Label	55.1	54.4	60.4	63.5	64.1	59.5	**69.3**	66.5	67.3	74.5	70.5	69.6
FixMatch	**55.4**	52.3	61.2	62.8	64.8	59.3	68.2	67.9	67.3	74.5	75.9	70.7
TCL	51.6	**56.6**	58.3	58.0	64.8	57.9	67.6	65.9	68.2	74.3	76.2	70.4
Ours	55.0	**56.6**	**63.9**	**65.3**	**67.5**	**61.7**	67.0	66.8	**69.9**	**77.1**	**79.1**	**72.0**

Table 3. **Seen Compositions**. When using $\geq$5% of the labeled training data our method outperforms outperforms semi-supervised baselines for recognition of adverbs in previously seen action-adverb compositions.

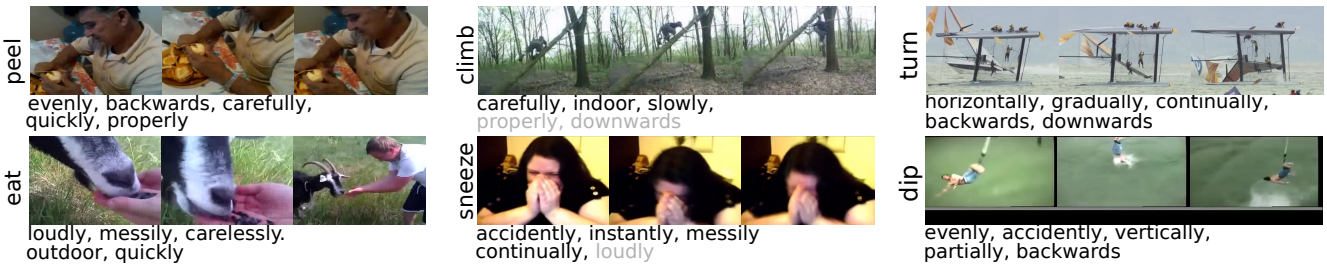

Figure 8. Example pseudo-labels from our proposed multi-adverb pseudo-labeling for the indicated action. Pseudo-labels below their adverb threshold are shown in grey. Our method can successfully identify multiple relevant adverbs for each video (left column), can use the adaptive thresholding to ignore incorrect or unnecessary pseudo-labels (middle), but can struggle to be accurate when actions co-occur (*downwards*, top right) and has no notion of situations where an adverb is infeasible (*backwards*, bottom right)

action-only data in each epoch of the labeled data. This allows better learning of the rarer adverb-action compositions. With 2 times the amount of action-only data the model is less likely to overfit to any noisy pseudo-labels early in training.

5.2. Task I: Seen Compositions

We test our method's suitability for recognizing adverbs in previously seen action-adverb compositions as in prior work [10]. Thus we use HowTo100M Adverbs [10] as well as VATEX Adverbs. For both datasets we test our approach with different amounts of labeled data used in training: 1%, 2%, 5%, 10% and 20%. The remaining training data is used in the action-only labeled set.

We compare to the **supervised-only** adverb recognition approach Action Modifiers [10] on which our approach is based. This learns from only adverb-labeled data. We also compare to several semi-supervised approaches: Pseudo-Label [27], FixMatch [58] and TCL [57], which we adapt for adverb recognition by combining them with Action Modifiers. This allows fair comparison as the backbone and adverb representations are the same in all methods. **Pseudo-Label** simply takes the most confident prediction of a data sample to be the pseudo label. **FixMatch** obtains pseudo-labels from weak augmentations of the input data. Strongly augmented versions are then trained to predict the same pseudo-label. This also uses fixed thresholding. Instead of the image augmentations used in FixMatch, **TCL** uses the video speed. It also optimizes agreement between the predictions for all classes rather than a single pseudo-label. Full implementation details can be found in supplementary.

Results are presented in Table 3. For VATEX Adverbs our approach outperforms or obtains competitive results over all baselines for all percentages of labeled data used. On HowTo100M Adverbs our approach outperforms baselines for the 5%, 10% and 20% labeled data settings. Our multi-adverb pseudo-labeling has more impact on VATEX Adverbs since it contains more adverbs. The improvement is also greater when using $\geq$ 5% labeled data. With fewer labels each adverb is seen in fewer situations, meaning the pseudo-labels become more noisy. We observe that TCL often performs worse than other approaches, despite being designed for video. This is because TCL encourages invariance to speed which affects adverbs such as *quickly* and *slowly*. Each of the semi-supervised baselines are comparable overall to the supervised-only method, this highlights the importance of our proposed multi-label pseudo-labelling and adaptive thresholding. Without these elements models are more biased to the particular action-adverb compositions.

We show examples of our multi-adverb pseudo-labeling in Fig. 8. Our method provides multiple relevant adverb pseudo-labels for each video. The model is able to use the adaptive thresholding to exclude incorrect predictions (climb *downwards*) or frequent compositions (climb *properly* and sneeze *loudly*). There are still noisy pseudo-labels such as climb *indoor* and dip *evenly*. There are also cases where the adverbs makes no sense in the context of the action, *e.g.* dip *backwards*. The incorrect prediction turn *downwards* highlights a challenge of adverb datasets, where there can be

Method	Accuracy
Supervised only	52.2
Ours	56.1
Training with full labels	65.1

Table 4. **Unseen compositions** in VATEX Adverbs. Our method improves generalization to unseen action-adverb compositions.

Method	MSR-VTT Adverbs	ActivityNet Adverbs
Source only	62.9	67.2
Pseudo-Label	63.9	66.4
Ours	65.0	66.6
Source + Target	67.5	71.6
Target only	70.5	71.8

Table 5. Transfer to **unseen domains** from VATEX-Adverbs. Our method aids generalization to similar domains (MSR-VTT Adverbs), but struggles with larger shifts (ActivityNet Adverbs).

multiple actions occurring at the same time. Here *downwards* refers to the people falling, rather than boat turning.

5.3. Task II: Unseen Compositions

We investigate whether our method can improve recognition of adverbs in previously unseen action-adverb pairs. We compare to supervised only Action Modifiers [10], although this was previously used with only seen pairs. Table 5 shows that ours improves the performance by 4%. The adaptive thresholding is key. Without it the pseudo-labels will primarily consist of previously seen adverbs compositions. However, there is much potential for future work as highlighted by the gap between ours and training with all compositions seen. Generalizing to unseen action-adverb combinations is necessary since it is infeasible to acquire sufficient labeled data for every possible action-adverb composition.

5.4. Task III: Unseen Domains

In Table 5 we test whether our pseudo-labeling approach can improve transfer to new domains. We compare our approach to training with only the source data, *i.e.* VATEX Adverbs, as well as the Pseudo-Label [27] baseline. Our method outperforms the Pseudo-Label approach for MSR-VTT Adverbs and gives a ∼2% gain over using only source domain videos. On ActivityNet Adverbs all three approaches are comparable, as the gap to this dataset is larger both in terms of action and adverb appearance and action length. Table 5 also shows the upper bounds when target data is used in training. The gap between our model's performance and source+target is relatively small, meaning adverb representations do transfer well between actions in different domains, however there is still much potential for improvement in the adverb representation itself. This is a more realistic setting to evaluate adverb representations, since labeled data is scarce. Transferring adverb representation to new domains is key to applications such as recognizing anomalous occurrences of an action or whether someone is following a recipe well.

Figure 9. We use the learned video-text embedding to identify zero-shot actions by compositions of adverbs and seen actions. We show each zero-shot action in bold alongside the closest action-adverb pair in the embedding space and one of the closest videos.

5.5. Describing Relationships Between Actions

We foresee many applications of adverbs in video understanding, such as detailed video captioning, describing anomalous instances of an action and feedback for people following instructions. Here we demonstrate qualitatively how adverbs can be used to identify fine-grained zero-shot actions by describing the relationship between these unseen actions and those previously seen. Fig. 9 shows examples of such actions. In each case the zero-shot action can be described by applying an adverb to a known action.

6. Discussion

Limitations. Our method has several limitations. Firstly, our model has no concept of infeasible combinations of action and adverbs and can be confounded by co-occuring actions where different adverbs apply. It also struggles when an adverb is labeled in few contexts. While our method can aid generalization to unseen action-adverb compositions and unseen domains, there is still far to go in these areas.

Potential Negative Impact. All datasets in this paper are sourced from YouTube and therefore the subjects and activities contained within are not representative of the diversity in global society. Thus our trained models will contain biases.

Conclusions. This paper has presented a semi-supervised method to recognize adverbs of actions. This allows us to understand how an action is being performed and understand fine-grained differences between actions. We propose multi-adverb pseudo-labeling to make use of videos with action-only labels. To cope with the long-tail distribution of adverbs and their action compositions our method also makes use of adaptive thresholding. We propose three new adverb recognition datasets which allow us to evaluate how well our method on recognizes adverbs in previously seen action-adverb compositions as well as unseen compositions and unseen domains. Results demonstrate our method improves performance in each of these three tasks.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. 2

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3365–3373, 2014. 2

[3] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018. 2

[4] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9922–9931, 2020. 2

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

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

[7] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10638–10647, 2020. 2

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2

[9] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9346–9355, 2019. 2

[10] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 868–878, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[11] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 919–929, 2020. 2

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

[13] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 214–229, 2020. 2

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5275, 2017. 4

[15] Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, and Cees GM Snoek. Motion-augmented self-training for video recognition at smaller scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10429–10438, 2021. 2

[16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5842–5850, 2017. 1, 2

[17] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances In Neural Information Processing Systems (NeurIPS)*, pages 529–536, 2004. 2

[18] Farnoosh Heidarivincheh, Majid Mirmehdi, and Dima Damen. Action completion: A temporal model for moment detection. In *British Machine Vision Conference (BMVC)*, 2018. 2

[19] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017. 2, 4

[20] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 709–727, 2020. 4

[21] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2758–2766, 2017. 4

[22] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10236–10247, 2020. 2

[23] Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. Videossl: Semi-supervised learning for video classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1110–1119, 2021. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5

[25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 4

[26] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Learning self-similarity in space and time as generalized motion for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13065–13075, 2021. 2

[27] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Internation Conference on Machine Learning (ICML) Workshops*, page 896, 2013. 2, 7, 8

[28] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2019. 4

[29] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2020. 2, 4

[30] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 4

[31] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. *arXiv preprint arXiv:2106.00168*, 2021. 4

[32] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 2

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

[34] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10900–10910, 2020. 4

[35] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3344, 2011. 2

[36] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C van Gemert. No frame left behind: Full video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14892–14901, 2021. 2

[37] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *British Machine Vision Conference (BMVC)*, 2019. 2

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[39] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, pages 1–19, 2015. 1

[40] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1049–1059, 2020. 2

[41] Pascal Mettes, William Thong, and Cees GM Snoek. Object priors for classifying and localizing unseen actions. *International Journal of Computer Vision (IJCV)*, 129(6):1954–1971, 2021. 2

[42] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2

[43] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. 4

[44] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval (ACMMM)*, pages 19–27, 2018. 2

[45] Augustus Odena. Semi-supervised learning with generative adversarial networks. *International Conference on Mahcine Learning (ICML) Workshops*, 2016. 2

[46] Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. *International Conference on Acoustics, Speech and Signal Processing*, 2021. 2, 4

[47] Bo Pang, Kaiwen Zha, and Cewu Lu. Human action adverb recognition: Adha dataset and a three-stream hybrid model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2325–2334, 2018. 1, 2

[48] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Differentiable grammars for videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11874–11881, 2020. 2

[49] AJ Piergiovanni and Michael S Ryoo. Fine-grained activity recognition in baseball videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1740–1748, 2018. 2

[50] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing (NeurIPS)*, 2015. 2

[51] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition (GCPR)*, pages 184–195, 2014. 4

[52] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt

Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision (IJCV)*, 119(3):346–373, 2016. 2

[53] Amir Rosenfeld and Shimon Ullman. Action classification via concepts and attributes. In *International Conference on Pattern Recognition (ICPR)*, pages 1499–1505, 2018. 2

[54] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[55] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2616–2625, 2020. 2

[56] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra-and inter-action understanding via temporal action parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 730–739, 2020. 2

[57] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10389–10399, 2021. 2, 7

[58] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 7

[59] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018. 2

[60] Jiahao Wang, Yunhong Wang, Sheng Liu, and Annan Li. Few-shot fine-grained action recognition via bidirectional attention and contrastive meta-learning. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, 2021. 2

[61] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4581–4591, 2019. 1, 2, 4

[62] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13214–13223, 2021. 2

[63] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 450–459, 2019. 2

[64] Huifen Xia and Yongzhao Zhan. A survey on temporal action localization. *IEEE Access*, 8:70477–70487, 2020. 1

[65] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. Multiview pseudo-labeling for semi-supervised learning from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[66] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 1, 2, 4

[67] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 2

[68] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 11562–11572, 2021. 2

[69] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 1

[70] Rowan Zellers and Yejin Choi. Zero-shot activity recognition with verb attribute induction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 2

[71] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1476–1485, 2019. 2

[72] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4486–4496, 2021. 2

[73] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 4

[74] Chen Zhu, Xiao Tan, Feng Zhou, Xiao Liu, Kaiyu Yue, Errui Ding, and Yi Ma. Fine-grained video categorization with redundancy reduction attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 136–152, 2018. 2

[75] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8746–8755, 2020. 2

[76] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020. 1

[77] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3537–3545, 2019. 2