EXPLORING SUB-PSEUDO LABELS FOR LEARNING FROM WEAKLY-LABELED WEB VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning visual knowledge from massive weakly-labeled web videos has attracted growing research interests thanks to the large corpus of easily accessible video data on the Internet. However, for video action recognition, the action of interest might only exist in arbitrary clips of untrimmed web videos, resulting in high label noises in the temporal space. To address this issue, we introduce a new method for pre-training video action recognition models using queried web videos. Instead of trying to filter out, we propose to convert the potential noises in these queried videos to useful supervision signals by defining the concept of Sub-Pseudo Label (SPL). Specifically, SPL spans out a new set of meaningful "middle ground" label space constructed by extrapolating the original weak labels during video querying and the prior knowledge distilled from a teacher model. Consequently, SPL provides enriched supervision for video models to learn better representations for downstream tasks. We validate the effectiveness of our method on four video action recognition datasets and a weakly-labeled image dataset to study the generalization ability. Experiments show that SPL outperforms several existing pre-training strategies using pseudo-labels and achieves competitive results on HMDB51 and UCF101 datasets compared with recent pre-training methods.

1 INTRODUCTION

Remarkable successes (Simonyan & Zisserman, 2014; Tran et al., 2015; Feichtenhofer et al., 2019) have been achieved in video recognition in recent years thanks to the development of deep learning models. However, training deep neural networks requires a large amount of human-annotated data. It requires tremendous human labor and huge financial cost and therefore oftentimes sets out to be the bottleneck for real-world video recognition applications.

Web videos are usually acquired by online querying through label keywords. A keyword, which we refer as a weak label, is then assigned to each untrimmed video obtained. Although large-scale videos with weak labels are easier to be collected, training with un-verified weak labels poses another challenge in developing robust models. Recent studies (Ghadiyaram et al., 2019; Kuehne et al., 2019; Chang et al., 2019) have demonstrated that, in addition to the label noise (e.g. incorrect action labels on untrimmed videos), there are temporal noise due to the lack of accurate temporal localization for the action. This means an untrimmed web video may include other non-targeted content or may only contain a small proportion of the target action. Reduce noise effects for large-scale weakly-supervised pre-training is critical but particularly challenging to be practical. (Ghadiyaram et al., 2019) suggests to query short videos (e.g., within 1 minute) to obtain more accurate temporal localization of target actions. However, such data pre-processing method prevents models from fully utilizing widely available web video data, especially longer videos with richer contents.

In this work, we propose a novel learning method to conduct effective pre-training on untrimmed videos from the web. Instead of simply filtering the potential temporal noise, we propose to convert such "noisy" data to useful supervision by leveraging the proposed concept of Sub-Pseudo Label (SPL). As shown in Figure 1, SPL creates a new set of meaningful "middle ground" pseudo-labels to expand the original weak label space. Our motivation is based on the observation that, within the same untrimmed video, these "noisy" video clips have semantic relations with the target action (weak label class), but may also include essential visual components of other actions (such as the teacher model predicted class). Our method aims to use the extrapolated SPLs from weak labels and distilled labels to capture the enriched supervision signals, encouraging learning better representations during pre-training that can be used for downstream fine-tuning tasks.

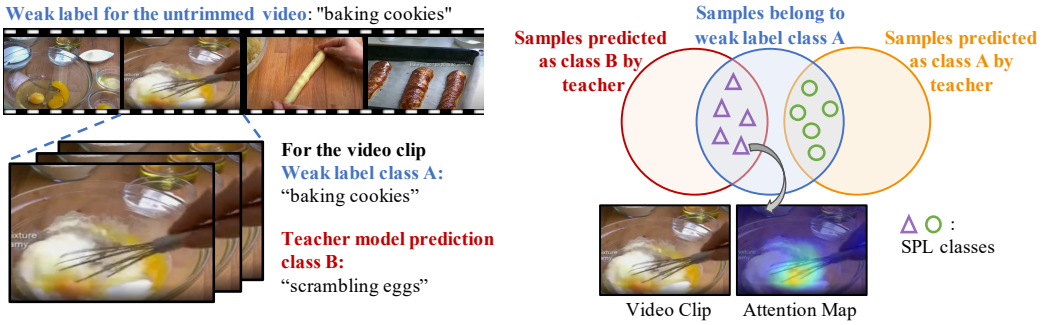

Figure 1: SPL converts the potential noises in untrimmed web videos to useful supervision signals by creating a new set of meaningful "middle ground" pseudo labels, such as mixing the eggs and flour shown in this example. Enriched supervision is provided for effective video model pre-training.

Discovering meaningful SPLs has critical impact on learning high quality representations. To this end, we take advantage of the original weak labels as well as the prior knowledge distilled from a set of target labeled data from human annotations. Specifically, we first train a teacher model from the target labeled data and perform inference on every clip of web videos. From the teacher model predicted labels and the original weak labels of the web video, we design a mapping function to construct SPLs for these video clips. We then perform large-scale pre-training on web videos utilizing SPLs as the supervision space. In addition, we study different variants of mapping functions to tackle high-dimensional label space when the number of classes is high. We construct weakly-labeled web video data based on two video datasets: Kinetics-200 (Xie et al., 2018; Carreira & Zisserman, 2017) and SoccerNet (Giancola et al., 2018). Experimental results show that our method can consistently improve the performance of conventional supervised methods and several existing pre-training strategies on these two datasets.

Our contributions can be concluded as follows: (a) We propose a novel concept of SPL to provide valid supervision for learning from weakly-labeled untrimmed web videos so that better representations can be obtained after the pre-training. (b) We investigate several space-reduced SPL classes discovering strategies utilizing weak labels as well as the knowledge distilled from the teacher model trained on the labeled dataset. (c) Experiments show that our method can consistently improve the performance of baselines on both common and fine-grained action recognition datasets. We also validate the generalization ability of the proposed method on a weakly-labeled image classification benchmark (the source code is provided).

2 Related Work

Learning from the web data. There are growing studies taking use of information from the Internet that aim to reduce the cost of data collection and annotations (Chen et al., 2013; Mahajan et al., 2018; Yan et al., 2019; Duan et al., 2020; Yalniz et al., 2019). For video classification, early works (Sun et al., 2015; Gan et al., 2016) focus on utilizing web action images to boost action recognition models, which do not consider the rich temporal dynamics of videos. Recently, (Ghadiyaram et al., 2019) demonstrates that better pre-training models can be obtained by learning from very large scale noisy web videos with short length e.g., within 1 minute. Instead, we propose SPL to handle the temporal noise in untrimmed videos and provide enriched supervision for pre-training.

Knowledge distillation. Our work is also related to Knowledge Distillation (Buciluǎ et al., 2006; Hinton et al., 2015) where the goal is to transfer knowledge from one model (the teacher) to another (the student). (Xie et al., 2020) shows that student models can outperform the teacher model based on unlabeled data and data augmentation. (Yan et al., 2019) finds that training student models using cluster assignments of features extracted from a teacher model can improve the generalization of visual representations. (Müller et al., 2020) proposes to improve the transfer by forcing the teacher to expand the network output logits. While to some extend SPL and it share similar high-level motivation of exploring sub-concepts, the differences are significant. Methods are different: our method uses the extrapolated SPLs from weak labels and distilled labels, while (Müller et al., 2020) learns to expend teacher network logits, which relies on a pre-defined hyper-parameter of subclass numbers. Problems are different: (Müller et al., 2020) focuses on the knowledge distillation between large and small networks for image classification.

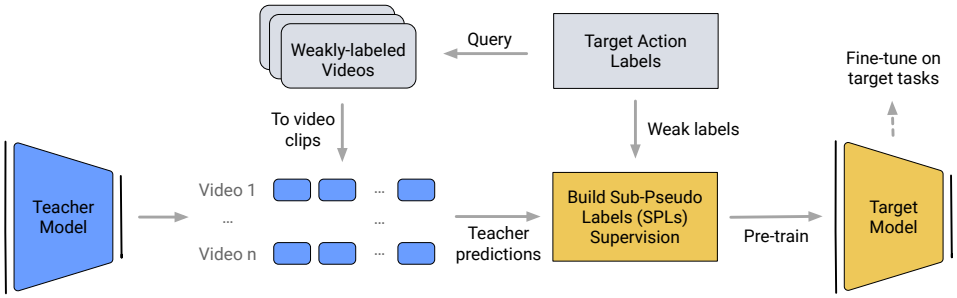

Figure 2: Pre-training framework for learning form web videos via exploring SPLs.

Video action recognition. State-of-the-art action recognition models either use two-stream (RGB and flow) networks (Simonyan & Zisserman, 2014; Feichtenhofer et al., 2016) or 3D ConvNets (Carreira & Zisserman, 2017; Tran et al., 2015; Feichtenhofer et al., 2019). The former one first uses 2D networks to extract image-level feature and then performs temporal aggregation on top while the latter one learns spatial-temporal features directly from video clips in the form of consecutive frames. Our work improves the performance of the action recognition model utilizing web videos.

3 METHOD

3.1 PROBLEM FORMULATION AND METHOD OVERVIEW

For pre-training on a dataset D_p with N target actions, we aim to learn representations that can benefit the downstream task by afterwards fine-tuning on the target dataset D_t. This pre-training process of model M is usually achieved by minimizing the cross-entropy loss between the data samples x and their corresponding labels y, as follows:

$$L_{\text{CE}} = -\mathbb{E}_{(x,y)\sim D_p} \sum_{c=1}^{N} y_c \log M(x), \tag{1}$$

where $y_c \in \{0, 1\}$ indicates whether x belongs to class $c \in [0, N-1]$.

In the case of pre-training on a web video set as D_p, we sample clips from these untrimmed web videos to construct the training data. Since there are no ground-truth annotations, assigning a valid label y for each clip sample x is a key. A common practice (Ghadiyaram et al., 2019; Mahajan et al., 2018) is to treat the text query or hash tags that come together with the web videos as weak labels l. However this causes high label and temporal noises as target actions might exist in arbitrary clips of the entire video that occupy a very small portion. In addition to relying on the weak labels, we can also distill knowledge from a teacher model T trained on the target dataset D_t using Eq. 1, where D_p is replaced by D_t. A basic teacher-student training pipeline (Furlanello et al., 2018; Xie et al., 2020) can be applied by treating the teacher model prediction as the pseudo-label to train a student model on D_p. But there will be information lost as the original informative weak labels are totally ignored. Another strategy is to use agreement filtering to select reliable data whose weak labels match their teacher model predictions. However, in practice we find this strategy will discard a large amount of training data from D_p (over 60% in our experiments on the Kinetics-200 dataset), which limits the data scale for training deep neural networks.

Instead of treating the potential noise in D_p as useless data to filter out, we propose to migrate such noisy data to useful supervision by defining the concept of Sub-Pseudo Label (SPL). Specifically, SPL creates a new set of meaningful "middle ground" pseudo-labels, which are discovered by taking advantage of the original weak labels and the prior knowledge distilled from the teacher model. Figure 2 illustrates the overall framework to utilize SPLs.

3.2 SPL FOR INDIVIDUAL TRAINING SAMPLE

To determine the SPL class for each video clip in D_p, we first perform inference on each video clip in D_p using the teacher model T trained on D_t. A 2-dimensional confusion matrix $C \in \mathbb{R}^{N\times N}$ can be obtained to summarize the alignments between the teacher model inferences (columns) and the original weak annotations (rows).

Specifically, video clips at the diagonal location (w, w) of C can be roughly treated as samples belonging to class w, which is agreed by the original weak label as well as the teacher model T.

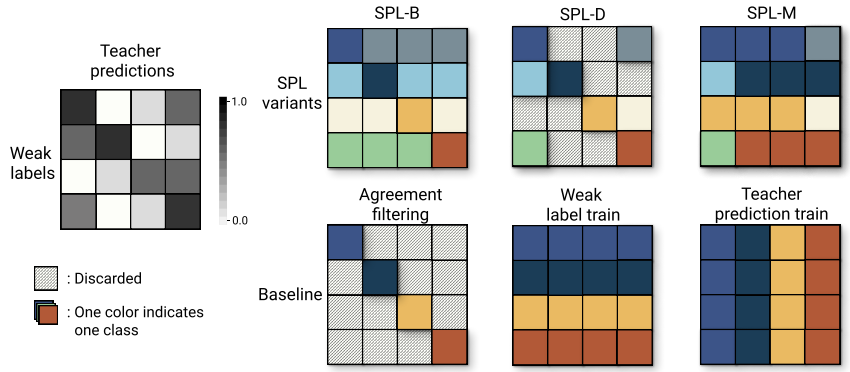

Figure 3: Illustration of the SPL mapping function. The left side is the confusion matrix between original weak labels and teacher predicted labels. The top-right illustrates three space-reduced variants of SPL. SPL-B: Using agreed and disagreed entries of each row as SPL classes. SPL-D: Using the top frequent entries as SPL classes. SPL-M: Merging less frequent off-diagonal entries to diagonals as SPL classes. The bottom-right illustrates three baseline pseudo-label strategies.

For other samples at off-diagonal location (h, w) of C, we interpret them as follows: from the view of the weak labels, these clips come from videos retrieved using text query of the action class h. Therefore, they include context information that may not exactly represent the action class h but is semantically related to it. However, from the view of the teacher model T, visual features that belong to action class w can also be found in these clips based on knowledge learned from the target dataset D_t. Instead of allocating these samples to ether action class h or w with the risk of leading to label noise, we convert such confusion to a useful supervision signal by assigning SPLs. For each data sample (x, y) in D_p, the sub-pseudo label $y \in [0, N^2 - 1]$ of the video clip x is obtained by:

$$y = N \cdot l + T(x), \tag{2}$$

where l is the weak label of x and $T(x)$ is its teacher prediction, where $l, T(x) \in [0, N - 1]$.

3.3 Reduction of the Quadratic SPL Space

Given N categories of original labels, SPL results in a quadratic label space $O(N^2)$. When N is big, the softmax output layer becomes too large to train 3D video models efficiently. Moreover, the distribution of SPL classes is usually long-tailed as some semantically distant classes can be unlikely confused with each other. We believe head SPL classes contain more useful supervision information than tails. Based on this motivation, we propose several variations of SPLs to reduce the SPL space.

Merge to Diagonal (SPL-M): Suppose we are targeting at SPLs with a total number of $K \times N$ classes, we kept N in-diagonal classes and then select the most frequent $(K - 1) \times N$ off-diagonal classes as new SPLs. For the samples of un-selected off-diagonal classes, we merge them into the diagonals of their corresponding rows. Since each row belongs to a class of weak labels, this strategy promotes original weak labels over teacher predictions.

Discard Tail Off-diagonal (SPL-D): The confusion matrix between weak labels and teacher predictions itself encodes information about label noises: the less frequent SPL classes have higher potentials to contain mislabeled data. Therefore, unlike SPL-M merging the un-selected classes to corresponding diagonals, SPL-D discards these training samples, resulting in a smaller yet potentially less noisy training dataset.

Binary Merge (SPL-B): We explore using agreement and disagreement between weak labels and teacher predictions of video clips as a criterion to reduce the SPL space. In this case, the confusion matrix entries are reduced to $2 \times N$ classes, including N in-diagonal classes (agreement) and N off-diagonal classes (disagreement) of each row.

All variants can be viewed as pruning the confusion matrix, as illustrated in Figure 3. Figure 3 also intuitively unifies other strategies discussed in Section 3.1. *Weak label train* is the weakly-supervised training studied by (Ghadiyaram et al., 2019). *Teacher prediction train* is a basic teacher-student distillation methods studied by (Furlanello et al., 2018; Xie et al., 2020). *Agreement filtering* only takes samples whose the weak label is matched with the teacher model prediction on it. We will investigate these alternatives in experiments.

4 EXPERIMENTS

4.1 TARGET DATASETS

We evaluate the proposed SPL algorithm on both common action recognition as well as fine-grained action recognition datasets. For the common action dataset, we mainly use Kinetics-200 (K200) (Xie et al., 2018) which is a subset of Kinetics-400 (Carreira & Zisserman, 2017). In total, it contains 200 action categories with around 77K videos for training and 5K videos for validation. Studies (Xie et al., 2018) show that evaluations on K200 can be well generalized to the full Kinetics. Due to the lack of computation resources required for extreme large-scale pretraining when taking full Kinetics as the target dataset, K200 results in an optimal choice for new algorithm explorations. We also conduct evaluations on popular HMDB51 (Kuehne et al., 2011) and UCF101 (Soomro et al., 2012) benchmarks following the standard protocol. For the fine-grained dataset, we conduct experiments on SoccerNet (Giancola et al., 2018) dataset, which is proposed for action event recognition in soccer broadcast videos. We use 5547 video clips for training and 5547 clips for validation obtained from different full-match videos. For the evaluation matrix, we focus more on the performance of classifying foreground action classes, which are sparsely occurred in the broadcast videos. Therefore, mean average precision without background class is adopted. We also discuss cases for dataset whose class names cannot be used as reliable search queries in Section A.5 of Appendix.

4.2 WEAKLY-LABELED DATA COLLECTION

To construct the pre-training dataset D_p for each target dataset, we collect untrimmed web videos retrieved by a text-based search engine similar to (Caba Heilbron et al., 2015; Chen & Gupta, 2015) and construct several dataset versions for following studies. Also see Appendix for more details.

WebK200. We treat the class names of Kinetics-200 dataset as the searching queries and use 4 languages for each query, including English, French, Spanish and Portuguese. We construct two web videos sets with different sizes: WebK200-147K-V with 147K videos and WebK200-285K-V with 285K videos (including more low-ranked videos returned by the search engine). We sample a number of video clips with the length of 10 seconds from the retrieved videos. The number of clip samples for each class is roughly balanced according to the practice in (Ghadiyaram et al., 2019).

WebS4. For the three foreground classes, we obtain the searching queries based on related terms from Wikipedia such as "free kick goal", "corner kick goal" resulting in 9 kinds of queries in total for these 3 foreground classes. For the Background class, we use "soccer full match" as the query. For each searching query, we use 3 languages including English, French and Spanish. We sample video clips with the length of 10 seconds and keep the number of clips for each class roughly balanced. Two web video sets are obtained with different number of total clips: WebS4-73K-C and WebS4-401K-C, where 73K and 401K represent the number of video clips in these two sets.

4.3 IMPLEMENTATION DETAILS

We use 3D ResNet-50 (Wang et al., 2018) with self-gating (Xie et al., 2018) as the baseline model and more details are described in the Appendix. Following (Wang et al., 2018), the network is initialized with ResNet-50 pre-trained on ImageNet (Deng et al., 2009). At training stage, we use the batch size of 6 and take 16 RGB frames with temporal stride 4 as the input. The spatial size of each frame is 224×224 pixels obtained from the same randomly cropping operation as (Wang et al., 2018). For the pre-training on our WebK200 sets, we set warm up training for 10 epochs with starting learning rate as 0.04 and then use learning rate 0.4 with cosine decay for 150 epochs. For the fine-tuning, we follow (Ghadiyaram et al., 2019) to initialize the model from the last epoch of the pre-training and conduct end-to-end fine-tuning. We set warm up training for 10 epochs with starting learning rate as 0.04 and then use learning rate of 0.4 with cosine decay for 60 epochs. For SoccerNet dataset, we use the same pre-training setting with WebK200 to conduct pre-training on the WebS4 sets. For the fine-tuning, we use learning rate of 0.005 for 20 epochs. More settings of hyper-parameters are described in the appendix. They are obtained to get the best performance of baselines. Our method is implemented using TensorFlow (Abadi et al., 2015).

4.4 RESULTS ON THE KINETICS-200 DATASET

In this section, we verify the effectiveness of the proposed method on the Kinetics-200 dataset via studies of different perspectives and explorations. Fine-tuning results are reported.

Table 1: Comparisons with different pre-training strategies on WebK200-147K-V set with 6.7×10^5 clips. Fine-tuning results on Kinetics-200 dataset are shown.

Pre-train Method	Top-1	Top-5
ImageNet Pre-train	80.6	94.7
Weak Label Train	82.8	95.6
Teacher Prediction Train	81.9	95.0
Agreement Filtering Train	82.9	95.4
Data Parameters	83.2	95.5
SPL-B (Ours)	**84.3**	**95.7**

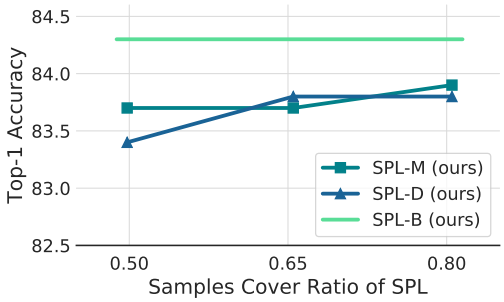

Figure 4: SPL-M and SPL-D with different samples cover ratio (SCR) defined in the Section 4.4

Comparisons with other pre-training strategies. Section 3.3 and Figure 3 categorize different pseudo-label strategies. Here we compare these strategies to ours. We report results based on their pre-training on our WebK200-147K-V set with 6.7×10^5 clips. From Table 1, we find they can all improve upon the baseline ImageNet pre-training. The performance gap between pre-training using Weak Label (Ghadiyaram et al., 2019) and Teacher Prediction (Xie et al., 2020) indicates there are more useful information included in weak labels. Although Agreement Filtering can do some noise reduction to the web videos, it discards around 60% of training samples resulting in a comparable performance with Weak Label. We also adopt Data Parameters (Saxena et al., 2019), one of the recent state-of-the-art methods for learning with noisy labels, to conduct pre-training on web videos. Our SPL-B (variation with the best performance on Kinetics-200) outperforms these baselines and is able to take use of all noisy data.

Comparisons between different variations of SPL. To compare the performance of the variants of SPL, we conduct experiments on our WebK200-147K-V set with 6.7×10^5 clips for pre-training. We start with total number of SPL classes as $K \times N = 400$ so that the label space is consistent for the three variations. The label space of SPL-D and SPL-B is controlled by hyper-parameter K and their space is reduced by merging or discarding samples belong to infrequent SPLs. There is a question about how many frequent SPLs to keep. More classes introduce more fine-grained tail SPLs yet higher computation cost. We define samples cover ratio (SCR) = $\frac{\text{\# of samples in selected SPLs}}{\text{\# of total samples}}$. Specifically, 400 SPL classes give SCR$= 49.81\%$. We evenly increase SCR by 15% to get 1600 and 4500 SPL classes with SCR of 65% and 80% respectively. From the result in Figure 4, we find that including more SPL classes can generally improve the performance of SPL-M and SPL-D. But the overall improvement gain is limited and they underperform SPL-M.

Effect of the number of training samples, more clips or more videos? It is a common practice to improve the performance of deep neural networks by increasing the number of training samples (Mahajan et al., 2018; Ghadiyaram et al., 2019). There are two practical options to do so given untrimmed web videos: (1) sampling more clips from a given number of web videos or (2) collecting more videos to sample clips. Both ways have potential drawbacks. The first one may result in duplication of visual contents. The second one may lead to lower quality of weak labels because this means we have to include more low-ranked videos returned by the search engine. In the following experiments, we aim to study this practical question and verify the effectiveness of SPL.

Effect of more clips. We sample different numbers of clips from WebK200-147K-V set (described in Section 4.2) and plot results of different pre-training strategies in Figure 5. The baseline result with red dot line represents the performance of using ImageNet pre-trained models. Results show that sampling more video clips from a given number of untrimmed videos can help improve the model performance. We also find that with a sufficient number of video clips available, our SPL methods consistently outperform weak label pre-training by providing rich and valid supervision knowledge.

Effect of more videos. We sample a similar number of total video clips, around 1.4×10^6, from WebK200-147K-V (147K videos) and WebK200-285K-V (285K videos) to obtain two training sets. We conduct teacher model inference on these two sets to get a synthetic measurement of the noise ratio and find this ratio is larger in WebK200-285K-V set. The comparison in Table 2 indicates that, though synthetic noise ratio is higher with the increase of videos, enriched visual contents are beneficial to some extent. More importantly, our SPL-B obtains more performance gain than directly using weak labels, which also indicates its robustness to label noise.

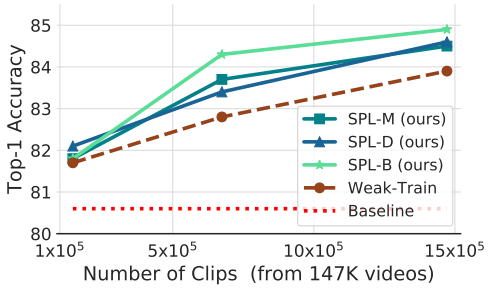

Figure 5: Effect of different numbers of clip given a fixed number of videos (WebK200-147K-V). Results are on Kinetics-200.

Table 2: Results on Kinetics-200 (Top-1 accuracy) based on pre-training with different numbers of videos for a similar number of total clips (around 1.4×10^6).

Pre-train Method	Number of videos	
	147K-V	285K-V
Synthetic Noise Ratio [1]	58.9 %	65.5%
Weak Label Train	83.9	84.0
SPL-M (Ours)	84.5	84.8
SPL-D (Ours)	84.6	84.9
SPL-B (Ours)	**84.9**	**85.3**

Comparisons with other methods on Kinetics-200. In Table 3, we list results of other existing methods on this benchmark: I3D (Carreira & Zisserman, 2017), S3D (Xie et al., 2018), R3D-NL (Wang et al., 2018), R3D-CGNL (Yue et al., 2018). The comparisons show that our method, which uses only extra 2-4x more web videos, is able to outperform the previous best reported number by a large margin (over 5%).

Comparisons with state-of-the-arts on HMDB51 and UCF101 datasets. We fine-tune the SPL-B pre-trained model from Webk200-285K-V on these two benchmarks to obtain results and also report our ImageNet initialized baselines. In Table 4, we find SPL improves our baseline significantly and also outperforms recent self and webly-supervised pre-training methods that relies on only visual modality: MemDPC (Han et al., 2020), SpeedNet (Benaim et al., 2020), CPD (Li & Wang, 2020), MIL-NCE (Miech et al., 2020), WVT (Stroud et al., 2020). Our method also uses a smaller number of videos compared with them. More complete comparisons are in Table 7 of Appendix, where we include results of more methods, settings of them and more discussions.

Table 3: Comparison with existing methods on Kinetics-200.

Method	Top-1	Top-5
I3D	78.0	-
S3D	78.4	-
R3D-50	75.5	92.2
R3D-50-NL	77.5	94.0
R3D-50-CGNL	78.8	94.4
R3D-101-NL	79.2	93.2
R3D-101-CGNL	79.9	93.4
SPL (Ours)	**85.3**	**96.6**

Table 4: Comparison with recent pre-training methods on HMDB and UCF (full version is in Table 7 of Appendix).

Method	Data	Model	HMDB	UCF
S3D-G	ImageNet	S3D-G	57.7	86.6
Our baseline	ImageNet	R3D-50-G	46.0	84.9
SpeedNet	K400 (240K)	S3D-G	48.8	81.1
MemDPC	K400 (240K)	R-2D3D	54.5	86.1
CPD	K400 (240K)	R3D-50	58.4	88.7
MIL-NCE	HT100M	S3D	61.0	91.3
WVT	WVT-70M	S3D-G	65.3	90.3
SPL (Ours)	WebK200-147K	R3D-50-G	**67.6**	**94.6**

Attention visualization of SPL classes. We visualize the visual concepts learned from SPLs. In Figure 6, we show some examples of SPL classes along with attention maps of the model trained using SPL. Attention maps are obtained using Grad-CAM (Selvaraju et al., 2017) to show the model's focus when making predictions. It is interesting to observe some meaningful "middle ground" concepts can be learnt by SPL, such as mixing the eggs and flour, using the abseiling equipment.

4.5 Experiments on SoccerNet Dataset

We also conduct experiments on SoccerNet (Giancola et al., 2018), a fine-grained action recognition dataset. Different from Kinetics-200 action classes, this dataset contains broadcast videos from soccer matches, so all classes contain sports actions sharing very similar visual (background) content. Therefore there exists high confusion between different classes. We use two web video sets WebS4-73K-C and WebS4-401K-C with 73K clips and 401K clips respectively as described in Section 4.2. Since the label space is not high, we use the full version of SPL that generates 4^2 SPL classes. In Table 5, we show the fine-tuning results on SoccerNet val set based on different types of pre-training. Our SPL method consistently outperforms other pre-training strategies.

[1] We use $\frac{\text{\# of off-diagonal elements}}{\text{\# of total elements}}$ in the confusion matrix as a synthetic measurement of the noise ratio.

Figure 6: Examples of attention visualization for SPL classes. Original weak label (blue) and the teacher model prediction (red) are listed. Some meaningful "middle ground" concepts can be learnt by SPL, such as mixing up the eggs and flour (top) and using the abseiling equipment (bottom).

Table 5: Results on the SoccerNet dataset. "Baseline" represents the ImageNet pre-training.

Method	Baseline	Weak Label	Agreement Filter	Teacher Pred.	SPL (Ours)
WebS4-73K-C pre-train	73.7	74.8	75.1	74.1	**76.1**
WebS4-401K-C pre-train	73.7	75.3	75.4	75.2	**76.8**

Table 6: Results on the Clothing1M dataset.

Method	None	Forward	CleanNet	NoiseRank	SPL (Ours)
Accuracy (%)	79.43	80.38	79.90	79.57	**80.50**

4.6 Generalizing SPL to Weakly-labeled Image Data

We study whether our proposed SPL has generalization ability to other domains other than videos. We test it on Clothing1M (Xiao et al., 2015), a large-scale image datasets with real-world label noises. Clothing1M contains $47,570$ training images with clean labels and $\sim$1M images with noisy labels. There are 14 original fashion classes. This dataset is challenging and 1% improvement is regarded as important. Since the label space is not high, we use the basic version of SPL that generates 14^2 SPLs for pre-training on the $\sim$1M images. We follow common experimental setting (Lee et al., 2018) and starts ResNet-50 pre-training with random initialization. Then we finetune on the clean set. Table 6 compares against previous methods such as None (Patrini et al., 2017), Forward (Patrini et al., 2017), CleanNet (Lee et al., 2018), NoiseRank (Sharma et al., 2020). SPL either outperforms or achieves competitive results.[2]

5 Conclusion

We propose a novel and particularly simple method of exploring SPLs from untrimmed web videos. Although previous literature has shown large-scale pre-training with weak labels can directly benefit, we demonstrate that SPLs can provide enriched supervision and bring much larger improvements. Importantly, SPL does not increase training complexity and can be applied in supervised, weakly or semi-supervised learning frameworks in orthogonal. We believe it is promising direction to discover meaningful visual concepts by bridging weak labels and the knowledge distilled from teacher models. SPL also demonstrates promising generalization to the image recognition domain, suggesting promising future directions like applying SPL to other tasks where there exists uncertainty in labels.

[2]We understand reproduce large-scale video training is costly. To encourage reproducibility of our method, we provide the source code of image recognition part in the supplementary material.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org. 5

Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv:1911.12667*, 2019. 12

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 *CVPR*, 2020. 7, 12

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, 2006. 2

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 5

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 3, 5, 7

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *CVPR*, 2019. 1

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv:2002.05709*, 2020. 12

Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015. 5

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 2

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. *arXiv:2003.13042*, 2020. 2

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 3

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1, 3

Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018. 3, 4

Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*, 2016. 2

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *CVPR Workshops*, 2018. 2, 5, 7

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 *ICCV*, 2017. 14

Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *ECCV*, 2020. 7, 12

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Workshop*, 2015. 2

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*, 2012. 12

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 12

Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019. 12

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 12

Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. Mining youtube-a dataset for learning fine-grained action concepts from webly supervised video data. *arXiv:1906.01012*, 2019. 1

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 5

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 12

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, 2018. 8

Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *arXiv:2001.05691*, 2020. 7, 12

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 2, 3, 6

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 7, 12

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. Subclass distillation. *arXiv:2002.03936*, 2020. 2

Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv:2003.04298*, 2020. 12

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017. 8

Serim Ryou, Seong-Gyun Jeong, and Pietro Perona. Anchor loss: Modulating loss scale based on prediction difficulty. In *ICCV*, 2019. 12

Shreyas Saxena, Oncel Tuzel, and Dennis DeCoste. Data parameters: A new family of parameters for learning a differentiable curriculum. In *NIPS*, 2019. 6

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 7

Karishma Sharma, Pinar Donmez, Enming Luo, Yan Liu, and I Zeki Yalniz. Noiserank: Unsupervised label noise reduction with dependence models. *arXiv:2003.06729*, 2020. 8

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1, 3

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 5

Jonathan C Stroud, David A Ross, Chen Sun, Jia Deng, Rahul Sukthankar, and Cordelia Schmid. Learning video representations from textual web supervision. *arXiv:2007.14937*, 2020. 7, 12

Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *ACM international conference on Multimedia*, 2015. 2

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv:1906.05743*, 2019. 12

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 3

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 5, 7, 12

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. In *CVPR*, 2018. 12

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 8

Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 2, 3, 4, 6

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 2, 5, 7, 12

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv:1905.00546*, 2019. 2

Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. *arXiv:1912.03330*, 2019. 2

Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact generalized non-local network. In *NIPS*, 2018. 7

A APPENDIX

A.1 MORE COMPLETE COMPARISONS ON HMDB51 AND UCF101 BENCHMARKS

We list a more complete comparisons with recent self and webly-supervised pre-training methods that relies on different modalities: OPN (Lee et al., 2017), ST-Puzzle (Kim et al., 2019), CBT (Sun et al., 2019), SpeedNet (Benaim et al., 2020), MemDPC (Han et al., 2020), CPD (Li & Wang, 2020), AVTS (Korbar et al., 2018), XDC (Alwassel et al., 2019), GDT (Patrick et al., 2020), MIL-NCE (Miech et al., 2020), WVT (Stroud et al., 2020). Compared with them, our SPL achieves competitive results by using a smaller number of videos to conduct the pre-training. This is also achieved by improving upon a weaker baseline model compared with R(2+1)D and S3D-G initialized by ImageNet pre-train weights.

Table 7: More complete comparisons on HMDB-51 and UCF-101 benchmarks. For modality, "V" refers to visual only, "A" represents audio and "T" means text description.

Method	Year	Data	Video num.	Model	Modality	HMDB	UCF
S3D-G	2018	ImageNet	-	S3D-G	V	57.7	86.6
R(2+1)D	2018	ImageNet	-	R(2+1)D	V	48.1	84.0
OPN	2017	UCF	13K	VGG	V	23.8	59.6
ST-Puzzle	2018	K400	240K	R3D-18	V	33.7	63.9
SpeedNet	2020	K400	240K	S3D-G	V	48.8	81.1
MemDPC	2020	K400	240K	R-2D3D	V	54.5	86.1
CPD	2020	K400	240K	R3D-50	V	58.4	88.7
AVTS	2018	Audioset	2M	MC3	V+A	61.6	89.0
XDC	2019	K400	240K	R(2+1)D	V+A	47.1	84.2
XDC	2019	IG65M	65M	R(2+1)D	V+A	67.4	94.2
GDT	2020	K400	240K	R(2+1)D	V+A	57.8	88.7
CBT	2019	K600	390K	S3D	V+T	44.5	79.5
MIL-NCE	2020	HT100M	1.2M	S3D	V+T	61.0	91.3
WVT	2020	WVT-70M	70M	S3D-G	V+T	65.3	90.3
Our baseline	-	ImageNet	-	R3D-50-G	V	46.0	84.9
SPL (Ours)	-	WebK200-147K	147K	R3D-50-G	V	**67.6**	**94.6**

A.2 MORE DETAILS OF NETWORK STRUCTURE AND IMPLEMENTATION

Network structure. We describe more details of our backbone network R3D-50-G which is based on ResNet50 (Wang et al., 2018). An illustration of the backbone can be found in Table 8. A feature gating module (Xie et al., 2018) is added after each residual block. Feature gating is a self attention mechanism that re-weights the channels based on context (e.g. the feature map averaged over time and space).

Training. For both pre-training and fine-tuning, we follow (Wang et al., 2018) to do random cropping on each input frame to get 224×224 pixels from a scaled video whose shorter side is 256. We also perform random horizontal flipping and photometric augmentations such as randomly adjust brightness, contrast, hue and saturation. Synchronous stochastic gradient descent (SGD) is applied to train the model. Following (Wang et al., 2018), we use a momentum of 0.9 and weight decay of 1×10^{-7}. Following (Chen et al., 2020; Wu et al., 2018), we add a linear projection head during the pre-training. We adopt dropout (Hinton et al., 2012) and trainable BatchNorm (Ioffe & Szegedy, 2015) with the same setting as (Wang et al., 2018). We also apply anchor loss (Ryou et al., 2019) for pre-training with emphasis on hard examples and find it sometimes can benefit the learning of SPLs. It brings around 0.3 improvements of top-1 accuracy on Mini-Kinetics-200 dataset when conducting SPL-B pre-training on WebK200-147K-V with 6.7×10^5 clips. But the benefit is not obvious when conducting SPL-B pre-training on WebK200-285K-V set as we can achieve 85.3 of top-1 accuracy without using the anchor loss. It also does not benefit SPL-M and SPL-D on Mini-Kinetics-200 dataset.

Inference. We follow (Wang et al., 2018) to perform inference on videos whose shorter side is re-scaled to 256. Following (Xie et al., 2018), we sample 64 frames from the whole videos and conduct inference on Mini-Kinetics-200 dataset. A stride of 4 is applied during this sampling process.

Block		Output sizes $T \times S^2 \times C$
input		$64 \times 224^2 \times 3$
conv_1	5×7^2 stride 1×2^2	$64 \times 112^2 \times 64$
pool_1	1×3^2 stride 1×2^2	$64 \times 56^2 \times 64$
res_2	$\begin{bmatrix} 3 \times 1^2 \\ 1 \times 3^2 \\ 1 \times 1^2 \end{bmatrix} \times 3$ feature gating	$64 \times 56^2 \times 256$
res_3	$\begin{bmatrix} t_i \times 1^2 \\ 1 \times 3^2 \\ 1 \times 1^2 \end{bmatrix} \times 4$ feature gating	$64 \times 28^2 \times 512$
res_4	$\begin{bmatrix} t_i \times 1^2 \\ 1 \times 3^2 \\ 1 \times 1^2 \end{bmatrix} \times 6$ feature gating	$64 \times 14^2 \times 1024$
res_5	$\begin{bmatrix} t_i \times 1^2 \\ 1 \times 3^2 \\ 1 \times 1^2 \end{bmatrix} \times 3$ feature gating	$64 \times 7^2 \times 2048$

Table 8: ResNet50-G architecture used in our experiments. The kernel dimensions are $T \times S^2$ where T is the temporal kernel size and S is the spatial size. The strides are denoted as temporal stride $\times$ spatial stride^2. For res_3, res_4, and res_5 blocks the temporal convolution only applies at every other cell. E.g., $t_i = 3$ when i is an odd number and $t_i = 1$ when i is even.

A.3 SPL AND CONFUSION MATRIX STATISTICS

When we discuss the different variations of SPL in the main paper, we mentioned the long-tailed propriety for fine-grained SPL classes, especially for the original SPL, SPL-M and SPL-D. To verify this, we calculate the number of samples for each SPL class for WebK200-285K-V with 6.7×10^5 clips. In Figure 7, we show the cases of SPL-B (left) and SPL-D (right) with 400 SPL classes. The first 200 classes are in-diagonal SPL classes and rest 200 are off-diagonal SPL classes. We can find the long-tailed propriety exists in the off-diagonal SPL classes of SPL-D (SPL-M has a similar case). We also find the situation is much better for SPL-B, where the distribution of number of samples is more even for in-diagonal and off-diagonal SPL classes.

In Figure 8, we also show confusion matrices on WebK200 and WebKS4. They are web videos collected for target datasets Mini-Kinetics-200 and SoccerNet respectively. This aims to illuminate the difference of web videos collected for the common action recognition dataset and the fine-grained action recognition dataset. We can observe there exists high confusion between different classes on fine-grained one.

A.4 MORE DETAILS OF WEB DATA COLLECTION

Here we include more details of web data collection described in Section 4.2 of the main paper. For data collection of WebK200 sets: WebK200-147K-V and WebK200-285K-V, we conduct duplicate checking on these two sets and web videos with addresses appearing in the validation set of Mini-Kinetics-200 are removed.

We also list the full query terms used in collection of WebS4 sets for the target dataset SoccerNet. For the three foreground classes in SoccerNet: Goal, Yellow/Red Card, Substitution, we obtain the searching queries based on related terms from Wikipedia resulting in 9 kinds of queries in total. The queries for these three foreground classes are as follows: “free kick goal”, “corner kick goal”, “own goal”, “overhead kick goal”, “last-minute goal”, “ghost goal”, “red card”, “yellow card”, “substitution of players”. For the Background class, it is hard to come up with queries for these background moments and people usually do not create video content to highlight them on the Internet. Therefore, we use “soccer full match” as the query and randomly sample clips from the retrieved full match games. Because for entire soccer games, moments of foreground classes are very rare and sparse

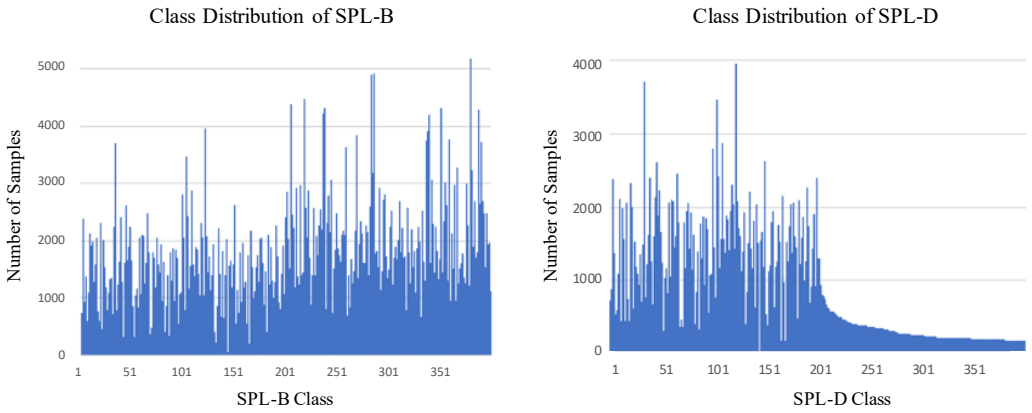

Figure 7: Class distribution of SPL on WebK200-285K-V set with 6.7×10^5 clips. We show the cases of SPL-B (left) and SPL-D (right) with 400 SPL classes in total. The first 200 classes are in-diagonal SPL classes and rest 200 are off-diagonal classes.

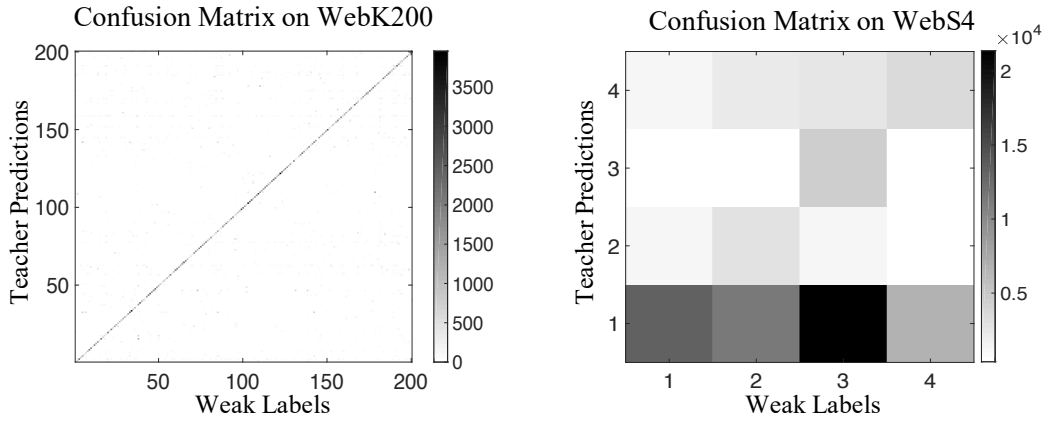

Figure 8: Confusion matrices on WebK200-285K-V (left, zoom in to better see off-diagonal elements) and WebS4-73K-C (right). They are web videos collected for target dataset Mini-Kinetics-200 and SoccerNet respectively. For WebS4, the four (1-4) classes are, Background, Yellow/Red Card, Goal, Substitution, respectively.

compared to the background events. Therefore, samples randomly sampled from the full game are very likely to belong to the Background class.

A.5 Research Challenges for Dataset that Has Special Class Names

One basic assumption for pre-training on web videos is that the search query is close related to the content of the retrieved videos so that these queries can be treated as weak labels to provide effective supervision during pre-training. There will be extra challenges when class names in the target dataset cannot be used as reliable search queries. For example, when conducting explorations on the fine-grained action datasets, we have considered the Something-Something (SS) (Goyal et al., 2017) dataset initially but are blocked due to this problem. We find that the quality of the collected web videos are quite low when using action classes from SS dataset as queries. The reason may be that the class names in SS are uncommon in the metda data of web videos resulting in the retrieved videos are almost unrelated to the search queries. Several examples for the retrieved videos are shown in Figure 9, Figure 10 and Figure 11 by randomly selecting class names in SS as the search query. In such a case, there would be extreme label noise if treating the text query as the class label for these web videos, which does not satisfy the basic assumption for learning from weakly-labeled

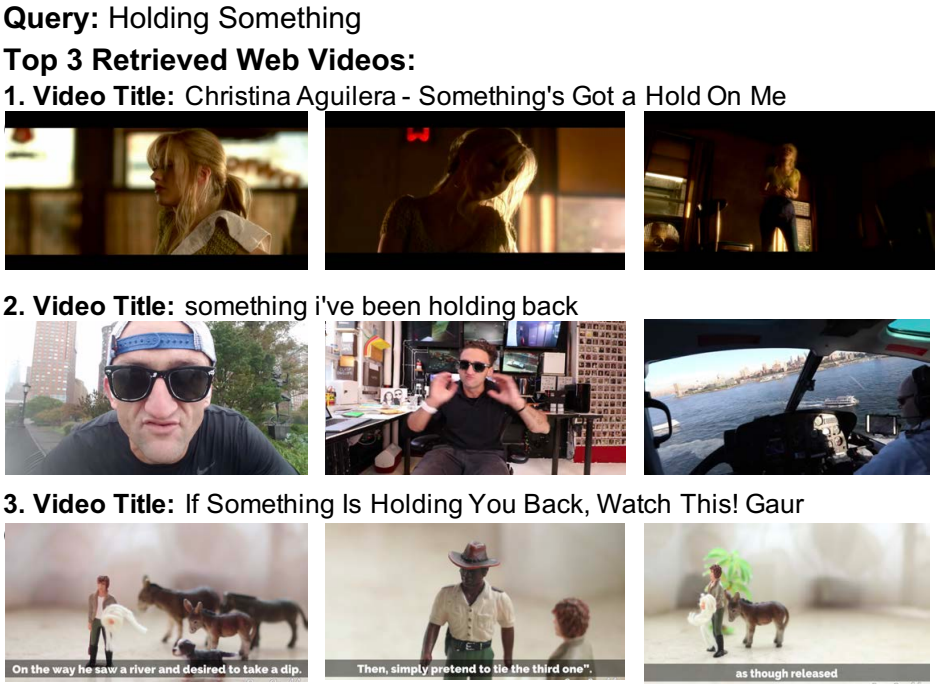

Figure 9: Top-3 retrieved videos obtained by taking the SS class "Holding Something" as the searching query. They are almost unrelated to the fine-grained motion defined in this class.

Figure 10: Top-3 retrieved videos obtained by taking the SS class "Closing Something" as the searching query. They are almost unrelated to the fine-grained motion defined in this class.

data. Addressing the acquisition of weakly-labeled web data with higher quality for this unique dataset could be a new research topic in this area.

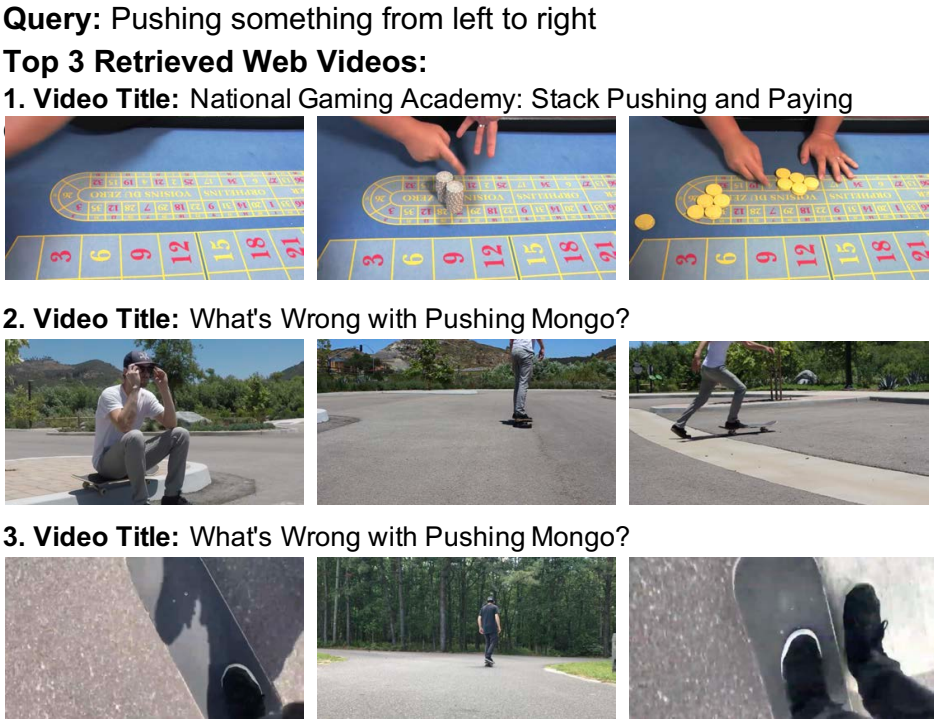

Figure 11: Top-3 retrieved videos obtained by taking the SS class "Pushing something from left to right" as the searching query. They are almost unrelated to the fine-grained motion defined in this class.

A.6 Source Code

To encourage reproducibility of our method in an reasonable amount of efforts, we provide the source code of image classification experiments, which can be found in the supplementary material, *SPL_code.zip*.