Large-scale weakly-supervised pre-training for video action recognition

Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, Dhruv Mahajan
Facebook AI
{deeptigp, mdf, trandu, xyan18, hengwang, dhruvm}@fb.com

Abstract

Current fully-supervised video datasets consist of only a few hundred thousand videos and fewer than a thousand domain-specific labels. This hinders the progress towards advanced video architectures. This paper presents an in-depth study of using large volumes of web videos for pre-training video models for the task of action recognition. Our primary empirical finding is that pre-training at a very large scale (over 65 *million videos), despite on noisy social-media videos and hashtags, substantially improves the state-of-the-art on three challenging public action recognition datasets. Further, we examine three questions in the construction of weakly-supervised video action datasets. First, given that actions involve interactions with objects, how should one construct a* verb-object *pre-training label space to benefit transfer learning the most? Second, frame-based models perform quite well on action recognition; is pre-training for good image features sufficient or is pre-training for spatio-temporal features valuable for optimal transfer learning? Finally, actions are generally less well-localized in long videos vs. short videos; since action labels are provided at a video level, how should one choose video clips for best performance, given some fixed budget of number or minutes of videos?*

1. Introduction

It is well-established [21, 33] that pre-training on large datasets followed by fine-tuning on target datasets boosts performance, especially when target datasets are small [3, 11, 27, 50]. Given the well-known complexities in constructing large-scale fully-supervised video datasets, it is intuitive that large-scale *weakly-supervised* pre-training is vital for video tasks.

Recent studies [37, 47, 63] have clearly demonstrated that pre-training on hundreds of millions (billions) of noisy web images significantly boosts the state-of-the-art in object classification. While one would certainly hope that successes would carry over from images [37, 47, 63] to videos, action recognition from videos presents certain unique challenges that are absent from the image tasks.

First, while web images primarily face the challenge of *label noise* (i.e., missing or incorrect object labels), for videos in the wild, the challenges are two-fold: label noise and *temporal noise* due to the lack of localization of action labels. In real-world videos, a given action typically occupies only a very small portion of a video. In stark contrast, a typical web image is a particular moment in time, carefully selected by its creator for maximal relevance and salience.

Second, in prior work on images, labels were restricted to scenes and objects (i.e., nouns). However, action labels (eg: “catching a fish”) are more complex, typically involving at least one verb-object pair. Further, even at large scale, many valid verb-object pairs may be observed rarely or never at all; for example, “catching a bagel” is an entirely plausible action, but rarely observed. Therefore, it is natural to inquire: is it more useful to pre-train on labels chosen from marginal distributions of nouns and verbs, do we need to pre-train on the observed portion of the joint distribution of (verb, noun) labels, or do we need to focus entirely on the target dataset’s labels? How many such labels are sufficient for effective pre-training and how diverse should they be?

Third, the temporal dimension raises several interesting questions. By analogy to images, short videos should be better temporally-localized than longer videos; we investigate this assumption and also ask how localization affects pre-training. In addition, longer videos contain more frames, but short videos presumably contain more relevant frames; what is the best choice of video lengths when constructing a pre-training dataset?

Finally, we question whether pre-training on videos (vs images) is even necessary. Both frame-based models and image-based pre-training methods like inflation [12] have been successful in action recognition. Is pre-training on video clips actually worth the increased compute, or, are strong image features sufficient?

In this work, we address all these questions in great detail. Our key aim is to improve the learned video feature representations by focusing exclusively on training data, which is complementary to model-architecture design. Specifically, we leverage over 65 million public, user-generated videos from a social media website and use the associated hashtags as labels for pre-training. The label noise and temporal noise makes our training framework *weakly-supervised*. Unlike all existing fully-supervised video datasets [36, 42, 43, 44, 61, 65] which required expensive annotation, our training data is truly extensible to billion-scale without incurring any annotation costs. We effectively tackle the aforementioned challenges with label space and temporal noise, and demonstrate significant performance gains on various target tasks. Overall, we summarize our findings:

- **Large-scale weak-supervision is extremely beneficial:** We show that large-scale video data, despite not providing strong supervision, tremendously helps models of varied capacities in learning better features. Our experiments clearly demonstrate that content diversity and scale outdo label and temporal noise.
- **Impact of data volume and model capacity:** We report interesting findings on the effect of pre-training data size, data sampling strategies, model capacities, etc. For instance, we find that increasing the training data (Sec. 4.1.1) improves performance while increasing model capacity exhibits interesting behavior (Sec. 4.1.2) .
- **What is a suitable pre-training video label space?** We systematically construct pre-training label sets that vary in cardinality and type (e.g., verbs, nouns, etc.), and study their effects of target tasks (Sec. 4.2). One key finding is that as in [47], pre-training labels that overlap the most with the target labels improve performance.
- **Short vs. long videos for pre-training?** We study the effect of pre-training on short vs. long videos (Sec. 4.3.1) and show that (a) for a fixed *video length budget* (e.g., 400K minutes of total training video duration), it is beneficial to choose a large number of short videos as they supply succinct localized actions compared to fewer long videos, (b) for a fixed *video budget* (e.g., 5M), choosing longer videos are beneficial as they offer diverse content.
- **Do we need pre-training on video data?** We investigate the true value of pre-training using video data. We show that it is necessary to use videos as opposed to video frames or images followed by inflation [12] to achieve better performance when operating at scale (Sec. 4.3.2).
- **State-of-the-art results:** We achieve a top-1 accuracy of **81.3%** on Kinetics, a 3.6% boost over the previous state-of-the-art [66] (Sec. 4.4). While the gains in [66] were achieved via architectural innovation, increased compute, etc., our boost is purely from pre-training a simple architecture (R(2+1)D [15]) at scale. On EPIC Kitchens action recognition challenge [17], we achieve an accuracy of **25.6%** on unseen (S2) test data, an improvement of 4.6% over the top entry in the leadership board at the time of submission. On Something-something [29], we achieve an accuracy of **51.6%**, a 2.1% improvement over the previous state-of-the-art [71].

2. Related Work

Learning from weak supervision: Given the known challenges in gathering exhaustive annotations for various image and video tasks, leveraging object labels and other meta information to supply weak supervision [8, 16, 20, 23, 31, 32, 40, 41, 51, 53, 54, 55, 56, 57, 60] is a widely-adopted approach. Orthogonal to this strategy, our work investigates transfer learning benefits when networks are pre-trained on weakly-supervised data, i.e., data afflicted with label and temporal noise. While novel changes to architectures have been proposed [58, 62] to combat label noise, our experiments demonstrate that large-scale training of an existing video architecture [15] makes it noise-resilient.

Dataset sources: Many prior approaches use Internet as a natural source of content [7, 13, 14, 19, 25, 26, 35, 38, 45, 47, 52, 59, 68] and the associated search queries, hashtags, or user-comments as labels to construct datasets. Most large-scale video datasets [2, 5, 39] were constructed by first curating a label taxonomy, analyzing the text metadata surrounding YouTube or Flickr videos, followed by some manual cleaning of the non-visual labels. [9, 22, 24, 48] analyzed movie scripts for automatic annotation of human actions for recognizing and localizing actions and actors. Our proposed work uses web-data and the associated text to supply weak supervision during pre-training.

Pre-training on large datasets: Since datasets for complex tasks such as object detection, segmentation and action recognition in videos are in smaller order of magnitude compared to ImageNet [18], pre-training on larger, auxiliary data followed by fine-tuning on target tasks [12, 21, 27, 37, 46, 47, 63, 15, 66] is very popular. Indeed, *inflation* [12] was proposed to exclusively leverage ImageNet instantiation by way of converting $2D$ filters to $3D$, given that training $3D$ models is computationally expensive. In this work, we show that pre-training on video clips performs significantly better than pre-training on image/video frames followed by inflation (Sec. 4.3.2).

3. Weak-supervision of video models

We harness millions of public videos from a social media website and use the associated hashtags as labels to supply weak supervisory signals while training video models. We construct and experiment with a variety of weakly-supervised datasets, which we describe next.

3.1. Source datasets

To construct pre-training video datasets, we use several *seed* action label sets and gather videos that are related to these labels. Specifically, for a given seed action label "catching a fish," we first construct all possible meaningful phrases (i.e., relevant hashtags) from it by taking the original and stemmed versions of every word in the seed label and concatenating them in all possible permutations. As an example, $\texttt{relevant\_hashtags}(\text{“}\textit{catching a fish}\text{”}) = \{\#catchingafish,\ \#catchfish,\ \#fishcatching, ...\}$. We then download public videos that are tagged with at least one of the hashtags from the set of $\texttt{relevant\_hashtags}(\text{“}\textit{catching a fish}\text{”})$ and associate them with the initial seed label. We use the seed labels as the final labels for videos during pre-training.

3.1.1 Properties of the source datasets

Seed labels: As we describe in Sec. 4.2, we study the effect of pre-training on different types of labels by considering four different seed label sets. The resulting source datasets are summarized in Table 1. We use a notation $\texttt{IG}-\texttt{source}-\texttt{size}$ throughout this paper, where *source* indicates the seed label set used and *size* indicates the number of videos[1]. Our primary training set uses 400 action labels from Kinetics [70] as seed labels, resulting in $\texttt{IG}-\texttt{Kinetics}$ dataset comprising 359 labels[2]. We also consider (a) the 1428 hashtags that match the 1000 synsets from ImageNet-1K [18], thereby constructing an $\texttt{IG}-\texttt{Noun}$ dataset[3], (b) the 438 verbs from Kinetics and VerbNet [64], thus an $\texttt{IG}-\texttt{Verb}$ dataset, and (c) all possible concatenations of the 438 verbs and the 1428 nouns from the above two seed label sets. We identify over $10,653$ such meaningful combinations[4], thus constructing an $\texttt{IG}-\texttt{Verb}+\texttt{Noun}$ dataset. More details on dataset construction are provided in the supplementary material. We want to reiterate that no manual annotation was involved in constructing these datasets implying that there is a large amount of noise in the data.

Long tail distribution: Distribution of videos in all our source datasets is heavily long-tailed. To mitigate its effect on model training, we adopt the square root sampling approach [49] when constructing pre-training data in our experiments as it proved to be most beneficial for images [47].

Diversity: Unlike all benchmark datasets which contain short videos of localized actions, video lengths in our source datasets range from 1 to 60 seconds. Thus, the labeled action can occur anywhere in a video, resulting in a large amount of temporal noise, an aspect we study in Sec. 4.3.

Given that we cannot make our datasets or the exact hashtags used public (as was the case with [63, 47]), we acknowledge that it is not possible for other research groups to reproduce our results at this time. Despite this limitation, we hope that the reader finds value in our wide-range of findings on various aspects of large-scale video pre-training that we describe in Sec. 4.

Pre-training dataset	Total #Videos	#Labels
IG-Kinetics	$65M$	359
IG-Noun	$19M$	1428
IG-Verb	$19M$	438
IG-Verb+Noun	$19M$	10653

Table 1. Statistics of the weakly-supervised datasets constructed for pre-training.

3.2. Target datasets

Next, we describe the target datasets used in experiments.

Kinetics [65]: Kinetics is a multi-class dataset with ~$246K$ training videos (400 human action labels). We report performance on the $20K$ validation videos.

EPIC-Kitchens [17]: EPIC-Kitchens is a multi-class egocentric dataset with ~$28K$ training videos associated with 352 noun and 125 verb classes. For our ablation studies, following [6] we construct a validation set of unseen kitchen environments. We evaluate our best pre-trained model on validation, standard seen (S1: 8047 videos), and unseen (S2: 2929 videos) kitchens test datasets.

Something-Something-v1 [29] is a multi-class dataset of ~$86K$ training videos and 174 fine-grained actions. We report results on the $11,522$ validation set.

Video deduplication: We devise a pipeline to deduplicate videos in the source datasets that may overlap with any from the target dataset. To err on the side of caution, we adopt an aggressive low-precision high-recall strategy and remove any potential duplicates (eg: we removed ~$29K$ videos from the IG-Kinetics dataset). Details are provided in the supplementary material.

3.3. Pre-training Setup

Models: R(2+1)D-d [15][5] is the fundamental architecture used for pre-training, where d denotes model depth = $\{18, 34, 101, 152\}$. As in [30], we construct models of depth $>$ 34 by replacing simple temporal blocks with bottleneck blocks for computational feasibility. We direct the reader to the supplementary material for details.

Loss Function: Our pre-training datasets are multi-label since multiple hashtags may be associated with any given video. The authors of [37, 47] have observed that per-label sigmoid outputs with logistic loss do not work well for noisy labels. Hence, we follow a simple strategy of randomly assigning one of the associated hashtags to each video thereby formulating a multi-class problem, and use softmax activations with cross-entropy loss.

Training Details: Video frames are down-sampled to a resolution of 128×171 and each video clip is generated by cropping a random patch of size 112×112 from a frame. Video clips of either 8 or 32 frames are used in our experiments, and temporal jittering is also applied to the input. Synchronous stochastic gradient descent (SGD) is used to

[1] We use $\texttt{IG}-\texttt{source}$ as notation whenever we refer to pre-training data source alone throughout this paper.

[2] For the remaining 41 labels, we could not find sufficient videos (i.e., at least 50 per label) using the approach detailed in Sec. 3.1.

[3] We get $1428 (> 1000)$ total hashtags because multiple hashtags may map to the same synset.

[4] Note that this is far fewer than $438 * 1428 =$~$600k$, as we discarded those concatenations which are not associated with at least 50 videos.

[5] Source code: https://github.com/dutran/R2Plus1D

train our models on 128 GPUs across 16 machines using caffe2 [10]. When 32 frames per input video clip are considered, each GPU processes 6 videos at a time (due to memory constraints), while 16 videos are processed at a time when 8 frames per video clip are considered. Batch normalization (BN) is applied to all convolutional layers and the statistics [34] are computed on each GPU. All pretraining experiments process $490M$ videos in total. Learning rate is set following the linear scaling procedure proposed in [28] with a warmup. An initial learning rate of 0.192 is used which is divided by 2 at equal steps such that the total number of learning rate reductions is 13 over the course of training.

4. Experiments

In this section, we study various aspects of large-scale weakly-supervised video pretraining. We first describe our evaluation setup and then report our extensive analysis on three aspects: (a) effect of scale, e.g., model capacity and pre-training data size, (b) design of the pre-training label space, and (c) temporal properties of videos. We also pretrain on benchmark datasets such as Sports-1M [39], Kinetics [65] as competitive baselines.

Evaluation Setup: As in [47], we consider two scenarios:

- **Full-finetuning (full-ft) approach** involves bootstrapping with a pre-trained model's weights and training end-to-end on a target dataset. We do a grid search for best the hyper-parameters (learning rate etc.) on validation data constructed by randomly holding out (10%) of training data. The hyper-parameters used for each experiment and target dataset are in the supplementary material. Full-ft approach has the disadvantage that it can potentially mask the absolute effect of pre-training for large target datasets.
- **Fully-connected (fc-only) approach** involves extracting features from the final fc layer of a pre-trained model and training a logistic regressor on each target dataset. This approach evaluates the strength of the learned features without changing the network parameters.

For multi-class target datasets our loss function is a $L2$-regularized logistic regressor and we report *accuracy*. For multi-label datasets, we use a per-label sigmoid output followed by logistic loss and report *mAP*. During testing, center crops of 10 clips uniformly sampled from each test video are considered, and the average of these 10 clip predictions are used to obtain the final video-level prediction.

4.1. Effect of large-scale

4.1.1 Amount of pre-training data

To understand this question, we pre-train on different amounts of training data by constructing different data subsets - IG-Kinetics- $\{500K, 1M, 5M, 10M, 19M, 65M\}$. R(2+1)D-34 models are independently trained on these data subsets on the exact same labels, with an input of 8-frames per video and evaluated on Kinetics (Fig. 1 (a)) and EPIC-Kitchens (Fig. 1 (b)).

As in [47, 63], we observe that performance improves log-linearly with training data size indicating that more pre-training data leads to better feature representations. For Kinetics, with full-ft approach, pre-training using $65M$ videos gives a significant boost of **7.8%** compared to training from scratch (74.8% vs. 67.0%). With increase in training data, performance gains are even more impressive when using fc-only approach, which achieves an accuracy of 73.0% with $65M$ training videos, thus closely matching the accuracy from full-ft approach (74.8%). On EPIC-Kitchens, using IG-Kinetics-65M yields an improvement of **3.8%** compared to using Kinetics for pre-training (16.1% vs. 12.3%). Compared with Kinetics, on EPIC-Kitchens, there is a larger gap in the performance between full-ft and fc-only approaches. This may be due to a significant domain difference in the pre-training and target label space.

These plots indicate that despite the dual challenge of label and temporal noise, pre-training using millions of web videos exhibit excellent transfer learning performance.

Data Sampling: Web data typically follows a Zipfian (long tail) distribution. When using only a subset of such data for pre-training, a natural question to ask is, if there are better ways to choose a data subset beyond random sampling. We design one such approach where we retain all videos from tail classes and only sub-sample head classes. We refer to this scheme as *tail-preserving* sampling.

Figure 1 (c) compares random and tail-preserving sampling strategies for Kinetics and reports performance obtained via fc-only approach. We observe that the tail-preserving strategy does consistently better and in fact, the performance saturates around $10M - 19M$ data points. Hence, for all future experiments, we adopted tail-preserving sampling strategy when needed.

4.1.2 Effect of model capacity

Table 2 reports the capacity of different video models and their effect on transfer learning performance. Specifically, we use IG-Kinetics-65M to pre-train 4 different R(2+1)D-d models, where $d = \{18, 34, 101, 152\}$ with input clip-length 32. On Kinetics, we observe that increasing model capacity improves the overall performance by 3.9%. In comparison, when training from scratch, the accuracy improves only by 2.7%. Interestingly, on EPIC-Kitchens, pre-training either using IG-Kinetics-65M or Kinetics (referred to as baseline) yield similar gains with the increase in model capacity. Unlike in [47] where the transfer learning performance was observed to be bottlenecked by capacity, we see a saturation in performance when going from $d = 101$ to $d = 152$[6]. Given that R(2+1)D-152 has higher GFLOPS

[6] For EPIC-Kitchens, we even observe a performance drop.

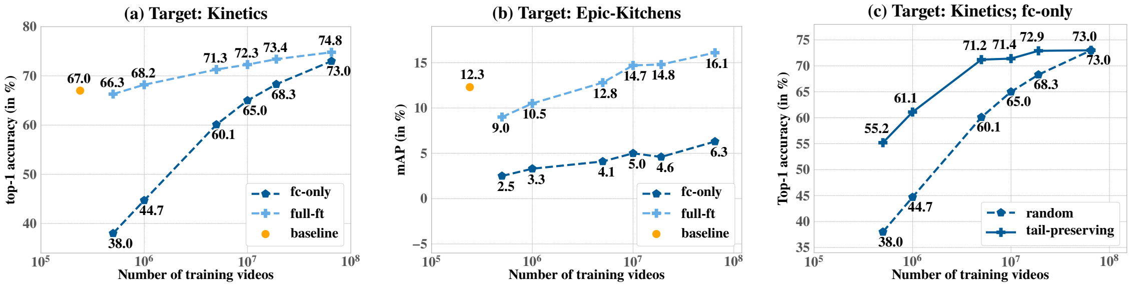

Figure 1. Illustrating the effect of increasing the number of pre-training videos. For Kinetics, we train a R(2+1)D-34 model from scratch as baseline, while for EPIC-Kitchens, we pre-train R(2+1)D-34 on Kinetics as baseline (indicated in orange). Random sampling was used for experiments reported in (a) and (b). X-axis is in log-scale.

			Kinetics		Epic-Kitchens	
Models	GFLOPS	# params	full-ft	baseline	full-ft	baseline
R(2+1)D-18	83	33M	76.0	69.3	20.8	14.8
R(2+1)D-34	152	64M	78.2	69.6	22.4	15.2
R(2+1)D-101	176	86M	79.1	71.7	24.9	17.1
R(2+1)D-152	252	118M	79.9	72.0	23.7	17.8

Table 2. Performance when pre-trained models of varied capacities are fully-finetuned on Kinetics (top-1 accuracy) and Epic-Kitchens (mAP). For EPIC-Kitchens, as a baseline, we use a model pre-trained on Kinetics.

compared to the largest image model in [47], we believe that our model may be bottlenecked by the amount of pre-training data. Thus, using more than $65M$ training videos may further boost the accuracy. Additionally, inability to do long-range temporal reasoning beyond 32 frames (due to memory constraints) may also be leading to this behavior. These questions are interesting to explore in the future.

4.2. Exploring the pre-training label space

Web videos and the associated (noisy) hashtags are available in abundance; hence it is natural to question: what constitutes a valuable pre-training label space for achieving superior transfer learning performance and how to construct one? Since hashtags are generally composed of nouns, verbs, or their combinations, and vary greatly in their frequency of occurrence, it is important to understand the trade-offs of different pre-training label properties (eg: cardinality and type) on transfer learning. In this section, we study these aspects in great detail.

4.2.1 Effect of the nature of pre-training labels

To study the type of pre-training labels that would help target tasks the most, as mentioned in Sec. 3.1, we systematically construct label sets that are verbs, nouns, and their combinations. Specifically, we use IG-Kinetics-19M, IG-Verb-19M, IG-Noun-19M, and IG-Verb+Noun-19M as our pre-training datasets. We use R(2+1)D-34 with clip-length of 32 for training. From Fig. 2, we may observe that for each target dataset, the source dataset whose labels overlap the most with it yield maximum performance. For instance, for Kinetics we see an improvement of *at least* 5.5%, when we use IG-Kinetics-19M for pre-training, compared to other pre-training datasets (Fig. 2(a)). Pre-training on IG-Noun benefits the noun prediction task of EPIC-Kitchens the most while IG-Verb significantly helps the verb prediction task (at least 1.2% in both cases, Fig. 2(b) and (c)). We found an overlap of 62% between IG-Verb and the verb labels and 42% between IG-Noun and the noun labels in EPIC-Kitchens. Pre-training on Sports-1M performs poorly across all target tasks, presumably due to its domain-specific labels.

Given that actions in EPIC-Kitchens are defined as verb-noun pairs, it is reasonable to expect that IG-Verb+Noun is the most well-suited pre-training label space for EPIC-Kitchens-actions task. Interestingly, we found that this was not the case (Fig. 2 (d)). To investigate this further, we plot the cumulative distributions of the number of videos per label for all four pre-training datasets in Fig. 3. We observe that though IG-Verb+Noun captures all plausible verb-noun combinations leading to a very large label space, it is also heavily skewed (and hence sparse) compared to other datasets. This skewness in the IG-Verb+Noun label space is perhaps offsetting its richness and diversity as well as the extent of its overlap with the EPIC-Kitchens action labels. Thus, for achieving maximum performance gains, it may be more effective to choose those pre-training labels that most overlap with the target label space while making sure that label distribution does not become too skewed. Understanding and exploiting the right trade-offs between these two factors is an interesting future research direction.

4.2.2 Effect of the number of pre-training labels

In Sec. 4.1.1, we study how varying the number of pre-training videos for a fixed source label space effects the transfer learning performance. In this section, we investigate the reverse scenario, i.e., vary the number of pre-training labels while keeping the number of videos fixed. We consider IG-Verb+Noun as our candidate pre-training dataset due to a large number $(10,653)$ of labels. We randomly[7] sub-sample different number of labels from the full label set all the way to 675 labels and fix the number of videos in each resulting dataset to be $1M$. We did not have enough training videos, i.e., at least $1M$ for fewer than 675 labels. Label sampling is done such that the smaller label

[7] Random sampling also makes sure that we remove uniformly from head and tail classes and long-tail issue with IG-Verb+Noun does not affect the observations.

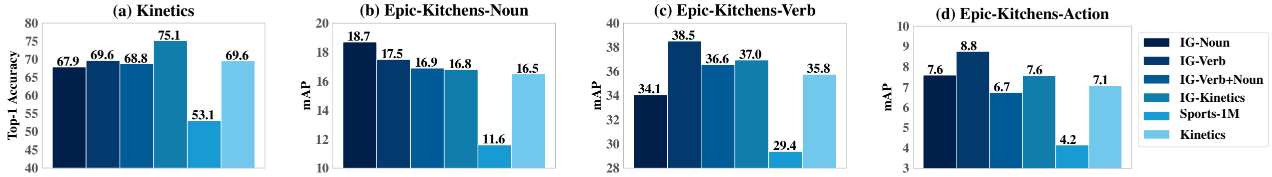

Figure 2. (a) Top-1 accuracy on Kinetics and (b)-(d) mAP on the three Epic-Kitchens tasks after fc-only finetuning, when different source label sets are used (indicated in the legend). The results indicate that target tasks benefit the most when their labels overlap with the source hashtags. Best viewed in color.

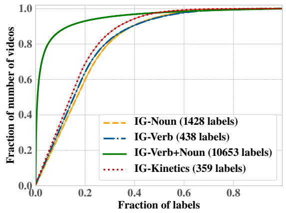

Figure 3. Cumulative distribution of the number of videos per label for the 4 pre-training datasets discussed in Sec. 4.2.1. The x-axis is normalized by the total number of labels for each dataset.

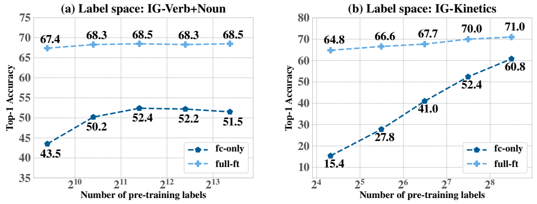

Figure 4. Top-1 accuracy on Kinetics when pre-training on different number of labels. Note that the source datasets used in panels (a) and (b) are different, hence the results are not comparable. X-axis is log scale.

space is a subset of the larger one. R(2+1)D-34 is used for pre-training with a clip-length of 8.

Figure 4 (a) shows performance on Kinetics. We may observe that using full-ft, there is an improvement of ~1% until 1350 labels, following which the performance saturates. For fc-only approach, the improvement in accuracy is ~9% before it saturates at 2700 labels. This suggests that the relatively fewer action labels in Kinetics (400) may not require a highly diverse and extensive pre-training label space such as IG-Verb+Noun. However, a large image label space ($17K$ hashtags) was proven [47] to be effective for highly diverse target image tasks (e.g., ImageNet-5k). Hence, we believe that to reap the full benefits of a large pre-training video label space, there is a need for more diverse benchmark video datasets with large label space.

Next, to understand the effect when the number of pre-training labels are fewer than the target labels (i.e, <400 for Kinetics), we consider IG-Kinetics as our pre-training dataset and vary the number of labels from 20 to 360. Pre-training data size is again fixed to $1M$. From Fig. 4 (b), we may observe a log-linear behavior as we vary the number of labels. There is a significant drop in the performance when using fewer labels even in the full-ft evaluation setting. This indicates that pre-training on a small label space that is a subset of the target label space hampers performance.

In summary, while using fewer pre-training labels hurts performance (Fig. 4 (b)), increasing the diversity through a simple approach of combining verbs and nouns (Fig. 4 (a)) does not improve performance either. Thus, this analysis highlights the challenges in label space engineering, especially for video tasks.

4.3. Exploring the temporal dimension of video

We now explore the temporal aspects of videos over long and short time scales. As mentioned in Sec. 3.1, our dataset inherently has large amounts of *temporal noise* as video lengths vary from 1 – 60 seconds and no manual cleaning was undertaken. While short videos are better localized, longer videos can potentially contain more diverse content. First, we attempt to understand this trade-off between temporal noise and visual diversity. Second, we address a more fundamental question of whether video clip-based pre-training is needed at all or is frame-based pre-training followed by inflation [12] is sufficient. The latter has an advantage of being very fast and more scalable.

4.3.1 Effect of temporal noise

To study this, we construct 3 datasets from IG-Kinetics:
(i) `short-N`: N videos of lengths between 1 – 5 seconds.
(ii) `long-N`: N videos of lengths between 55 – 60 seconds.
(iii) `long-center-N`: N videos (4 second long) constructed from the center portions of videos from `long-N`.
We ensure that the temporal dimension is the only factor that varies by keeping the label space and distribution (videos per label) fixed across these 3 datasets. Temporal jittering is performed for all these datasets during pre-training. Also, note that the exact same number of videos are seen while training on all the datasets. We now consider the following two scenarios.

Fixed number of videos budget (F1): A natural question that arises is: given a fixed budget of videos, what temporal property should guide our selection of pre-training videos? To answer this, we fix the total number of unique videos to $5M$ and consider `short-5M`, `long-5M`, and `long-center-5M` datasets. Note that both `short-5M` and `long-center-5M` have similar per-video duration (i.e., 4 seconds on average), but `long-center-5M` has greater temporal noise, since short videos are presumably more temporally localized than any given portion of longer videos. Between `short-5M` and `long-5M`, while `short-5M` has better temporal localization, `long-5M` may have greater content diversity From Table 3, we may

	`long-5M`	`long-500K`	`short-5M`	`long-center-5M`
F1	60.6	-	57.4	51.4
F2	-	50.6		

Table 3. Video top-1 accuracy when R(2+1)D-34 is pre-trained on 4 different short and long video datasets, followed by fc-only finetuning on Kinetics.

observe that `short-5M` performs significantly better than `long-center-5M` suggesting that short videos do provide better temporal localization. Also, `long-5M` performs better than `short-5M` by 3.2% indicating that more diverse content in longer videos can mask the effect of temporal noise. Thus, for a fixed total number of videos, longer videos may benefit transfer learning than short videos.

Fixed video time budget (F2): If storage or bandwidth is a concern, it is more practical to fix the total duration of videos, instead of the total number. Given this fixed budget of video hours, should we choose short or long videos? To answer this, we consider `short-5M`, `long-center-5M` and `long-500K` datasets, all with similar total video hours. From Table 3, we observe that `short-5M` significantly outperforms `long-500K`. This indicates that diversity and/or temporal localization introduced by using more short videos is more beneficial than the diversity within fewer long videos. Thus, for a fixed video duration budget, choosing more short videos yields better results. `long-center-5M` and `long-500K` perform similarly, indicating that on average, a fixed central crop from a long video contains similar information to a random crop from a long video. `short-5M` outperforms `long-center-5M`, consistent with the claim that short videos do indeed have better temporal localization.

4.3.2 Frame- vs. clip-based pre-training:

Although we have shown substantial gains when using clip-based R(2+1)D models for large-scale weakly supervised pre-training, it is computationally more intensive than $2D$ (image) models. Moreover, techniques such as *inflation* [12] efficiently leverage pre-trained image models by converting $2D$ filters to $3D$ and achieve top-performance on benchmark datasets. Given these, we want to understand the key value in pre-training directly on weakly-supervised video clips vs. images.

Towards this end, we first construct an image variant of the IG-Kinetics dataset (suffixed by $-Images$ in Table 4), following the procedure described in Sec. 3.1. We pre-train an 18 layer $2D$ deep residual model ($R2D$) [30] from scratch on different types of $2D$ data (image/single video frames). We then *inflate* [12] this model to $R3D$[8] [15] and perform full-finetuning with a clip-length of 8 on Kinetics.

From the inflation-based models in Table 4, we may observe that, pre-training on ImageNet achieves an improvement of 0.9% compared to training $R3D$ from scratch, while pre-training on IG-Kinetics-19M-Images yields a modest boost of 0.5% over ImageNet. Training on random video frames from IG-Kinetics-19M gives a further improvement of 0.5% over weakly-supervised image pre-training and an overall boost of 1.0% over ImageNet. To make sure that this marginal improvement is not because of pre-training on only $19M$ weakly-supervised noisy images, we pre-train using IG-Kinetics-250M-Images but find no further improvements. Finally, pre-training $R3D$ directly using video clips achieves an accuracy of 71.7%, a significant jump of 4.2% over the best inflated model (67.5%). This clearly indicates that effectively modeling the temporal structure of videos in a very large-scale pre-training setup is extremely beneficial.

Input dataset	Pre-training Input	Pre-train model	FT model	Top-1
ImageNet	Image	R2D-18	R3D-18	66.5
IG-Kinetics-19M-Images	Image	R2D-18	R3D-18	67.0
IG-Kinetics-250M-Images	Image	R2D-18	R3D-18	67.0
IG-Kinetics-19M	Video frame	R2D-18	R3D-18	67.5
Kinetics	Video clip	R3D-18	R3D-18	65.6
IG-Kinetics-19M	Video clip	R3D-18	R3D-18	**71.7**

Table 4. Understanding the benefit of using images vs. videos for pre-training.

4.4. Comparisons with state-of-the-art

In this section, we compare R(2+1)D-34 and R(2+1)D-152 models pre-trained on IG-Kinetics-65M with several state-of-the-art approaches on 3 different target datasets. For the results reported in this section alone, we follow [12] to perform fully-convolutional prediction for a closely-fair comparison with other approaches. Specifically, the fully-connected layer in R(2+1)D is transformed into a $1 \times 1 \times 1$ convolutional layer (while retaining learned weights), to allow fully-convolutional evaluation. Each test video is scaled to 128×171, then cropped to 128×128 (a full center crop). We also report results from using another frame scaling approach (indicated as SE in Tables 5 - 7), where each (train / test) video's shortest edge is scaled to 128, while maintaining its original aspect ratio, followed by a full center crop.

We note that each approach being compared varies greatly in terms of model architectures, pre-training datasets (ImageNet vs. Sports-1M), amount and type of input data (RGB vs flow vs audio, etc.), input clip size, input frame size, evaluation strategy, and so on. We also note that many prior state-of-the-art models use complex, optimized network architectures compared to ours. Despite these differences, our approach of pre-training on tens of millions of videos outperforms all existing methods by a substantial margin of **3.6%** when fully-finetuned on Kinetics (Table 5). Further, instead of uniformly sampling 10 clips, we used SC-Sampler [4] and sampled 10 salient clips from test videos and achieved a top-1 accuracy of **82.8%**.

In Table 6, we report the performance on the validation [6], seen (S1), and unseen (S2) test datasets that are part of the EPIC-Kitchens Action Recognition Challenge [1].

[8] We chose to inflate to $R3D$ because it was not immediately obvious how to inflate a $2D$ model to R(2+1)D given that it factorizes $3D$ convolution to $2D$ spatial and $1D$ temporal [15].

Method; pre-training	top-1	top-5	Input type
I3D-Two-Stream [12]; ImageNet	75.7	92.0	RGB + flow
R(2+1)D-Two-Stream [15]; Sports-1M	75.4	91.9	RGB + flow
3-stream SATT [70]; ImageNet	77.7	93.2	RGB + flow + audio
NL I3D [66]; ImageNet	77.7	93.3	RGB
R(2+1)D-34; Sports-1M	71.7	90.5	RGB
Ours R(2+1)D-34; IG-Kinetics	79.1	93.9	RGB
Ours R(2+1)D-34; IG-Kinetics; SE	79.6	94.2	RGB
Ours R(2+1)D-152; IG-Kinetics	80.5	94.6	RGB
Ours R(2+1)D-152; IG-Kinetics; SE	81.3	95.1	RGB
Ours R(2+1)D-152 + SC-Sampler [4]; IG-Kinetics; SE	**82.8**	**95.3**	RGB

Table 5. Comparison with the state-of-the-art on Kinetics. SE: short edge scaling.

Since the training dataset of EPIC-Kitchens consists of only ~$20K$ videos, for stronger baselines, we pre-train separate R(2+1)D-34 models on Kinetics and Sports-1M and fine-tune on EPIC-Kitchens. We also report the top-performing method from the challenge website [1] at the time of this manuscript submission. From Table 6, we may observe that, on unseen kitchens (S2), R(2+1)D-152 pre-trained on IG-Kinetics-65M improves the top-1 accuracy on verbs and nouns by **8.9%** and **9.1%** compared to R(2+1)D-34 pre-trained on Kinetics; and a **7.3%** boost on actions compared to R(2+1)D-34 pre-trained on Sports-1M. Similar substantial gains hold for seen (S1) and validation datasets. We note that we process only 32 RGB frames of the input video (no optical flow), at a much lower resolution (128 × 128) compared to the state-of-the-art, which is an ensemble model.

Finally, we report the performance (Table 7) on the validation data of Something-V1 [29], a challenging dataset with fine-grained classes. Using only RGB as input, pre-training with IG-Kinetics-65M achieves a top-1 accuracy of 51.6%, an improvement of **2.1%** over state-of-the-art [71][9] (49.5%). Compared to other approaches that use only RGB as input [69], our approach yields a boost of 3.4%.

5. Discussion

In this work, we explored the feasibility of large-scale, noisy, weakly-supervised pre-training with tens of million of videos. Despite the presence of significant noise in label space and temporal localization, our pre-trained models learn very strong feature representations. These models are able to significantly improve the state-of-the-art action recognition results on the popular Kinetics [65], a recently introduced EPIC-Kitchens [17], and Something-something [29] datasets. All of our large-scale pre-trained models show significant gains over Kinetics and Sports-1M, the de facto pre-training datasets in the literature. Our ablation studies address many important questions related to scale, label space, and temporal dimension, while also raising other interesting questions.

Method; pre-training	Verbs		Nouns		Actions	
	top-1	top-5	top-1	top-5	top-1	top-5
Test Unseen (S2)						
Leaderboard [1]	54.5	**81.2**	30.4	55.7	21.0	39.4
R(2+1)D [15]						
d=34; Kinetics	48.4	77.2	26.6	50.4	16.8	31.2
d=34; Sports-1M	47.2	77.4	28.7	50.0	18.3	31.6
Ours: d=34; IG-Kin.	55.5	80.9	33.6	56.7	23.7	39.1
Ours: d=34; IG-Kin. ; SE	56.0	80.6	32.4	55.6	23.6	39.5
Ours: d=152; IG-Kin.	55.3	80.3	34.7	58.2	25.4	40.7
Ours: d=152; IG-Kin.; SE	**57.3**	81.1	**35.7**	**58.7**	**25.6**	**42.7**
Test Seen (S1)						
Leaderboard [1]	**66.1**	**91.3**	**47.9**	**72.8**	**36.7**	**58.6**
R(2+1)D [15]						
d=34; Kinetics	59.1	87.4	38.0	62.7	26.8	46.1
d-34; Sports-1M	59.6	87.2	43.7	67.0	31.0	50.3
Ours: d=34; IG-Kin.	63.3	87.5	46.3	69.6	34.4	54.2
Ours: d=34; IG-Kin. ; SE	63.2	87.6	45.4	68.7	33.4	52.4
Ours: d=152; IG-Kin.	63.8	87.7	45.3	68.3	34.1	53.5
Ours: d=152; IG-Kin.; SE	65.2	87.4	45.1	67.8	34.5	53.8
Validation						
Baradel *et. al.* [6]; -	40.9	-	-	-	-	-
R(2+1)D [15]						
d=34; Kinetics	46.8	79.2	25.6	47.5	15.3	29.4
d=34; Sports-1M	50.0	79.8	24.8	46.2	16.0	30.3
Ours: d=34; IG-Kin.	56.6	83.5	32.7	55.5	22.5	39.2
Ours: d=34; IG-Kin. ; SE	55.5	83.3	34.8	57.2	22.8	39.8
Ours: d=152; IG-Kin.	56.6	83.8	34.5	58.5	23.5	40.6
Ours: d=152; IG-Kin.; SE	**58.4**	**84.1**	**36.9**	**60.3**	**26.1**	**42.7**

Table 6. Comparison with the state-of-the-art approaches on Epic-Kitchens dataset. IG-Kin. refers to IG-Kinetics. SE: short edge scaling.

Method; pre-training	top-1	top-5	Input type
NL I3D + Joint GCN [67]	46.1	76.8	RGB
S3D-G [69]	48.2	78.7	RGB
ECO_{En}Lite [71]	46.4	-	RGB
ECO_{En}Lite [71]	49.5	-	RGB + flow
R(2+1)D-34; Kinetics	45.2	74.1	RGB
R(2+1)D-34; Sports-1M	45.7	74.5	RGB
Ours: R(2+1)D-34; IG-Kin.	49.7	77.5	RGB
Ours: R(2+1)D-34; IG-Kin.; SE	49.9	77.5	RGB
Ours: R(2+1)D-152; IG-Kin.	51.0	**79.0**	RGB
Ours: R(2+1)D-152; IG-Kin.; SE	**51.6**	78.8	RGB

Table 7. Comparison with the state-of-the-art on Something-V1 [29]. IG-Kin. refers to IG-Kinetics. SE: short edge scaling.

Our study of label spaces found that sampling from the joint distribution of verb-noun pairs performs relatively poorly; this is presumably due to the skewed distribution and suggests that solving low-shot learning at scale is an important area of investment. Data augmentation can also help here: social media offers a nearly unlimited supply of unlabeled video, and non-parametric approaches like K-nearest neighbors could be employed to enhance sparse labels. Optimizing label space granularity in the face of data sparsity is another worth-while direction; this may require finding algorithmic ways to construct hashtag taxonomies for videos. Future research should also invest in creating new public benchmarks with larger and richer label spaces. Label spaces of current target datasets are small; they do not reflect the value of large-scale pre-training.

Finally, our analysis of temporal localization raises more questions. Our experiments clearly show the competing benefits of both good temporal localization in short videos, and greater diversity in long videos. Understanding this trade-off more rigorously may lead to intelligent data construction strategies that can leverage the best of both worlds.

[9] This number is achieved using RGB+flow and an ensemble of models.

References

[1] EPIC-Kitchens Action Recognition Challenge. [Online] Available https://competitions.codalab.org/competitions/20115#results. 7, 8

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[3] M. Andriluka, U. Iqbal, A. Milan, E. Insafutdinov, L. Pishchulin, J. Gall, and B. Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 1

[4] B. Korbar, D. Tran, L. Torresani. SCSampler: Sampling salient clips from video for efficient action recognition. *arXiv preprint arXiv:1904.04289*, 2019. 7, 8

[5] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. The new data and new challenges in multimedia research. *CoRR, abs/1503.01817*, 2015. 2

[6] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori. Object level visual reasoning in videos. *arXiv preprint arXiv:1806.06157*, 2018. 3, 7, 8

[7] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, 2010. 2

[8] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 2

[9] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *ICCV*, 2013. 2

[10] Caffe2 Team. Caffe2 : A new lightweight, modular, and scalable deep learning framework. [Online] Available https://caffe2.ai/. 4

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016. 1

[12] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *CVPR*, 2017. 1, 2, 6, 7, 8

[13] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015. 2

[14] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 2

[15] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. *CVPR*, 2018. 2, 3, 7, 8

[16] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 2

[17] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2, 3, 8

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009. 2, 3

[19] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 2

[20] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 2012. 2

[21] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *ICML*, 2014. 1, 2

[22] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *CVPR*. IEEE, 2009. 2

[23] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*, 2017. 2

[24] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is buffy–automatic naming of characters in tv video. 2006. 2

[25] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010. 2

[26] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from internet image searches. *Proceedings of the IEEE*, 2010. 2

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 2

[28] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 4

[29] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 2, 3, 8

[30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 7

[31] S. Hong, J. Oh, H. Lee, and B. Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*, 2016. 2

[32] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. *arXiv preprint arXiv:1701.00352*, 2017. 2

[33] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 1

[34] S. C. Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015. 4

[35] H. Izadinia, B. C. Russell, A. Farhadi, M. D. Hoffman, and A. Hertzmann. Deep classifiers from image tags in the wild. In *Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*, 2015. 2

[36] Y. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. 2014. http://crcv.ucf.edu/THUMOS14. 2

[37] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. *ECCV*, 2016. 1, 2, 3

[38] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016. 2

[39] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. *CVPR*, 2014. 2, 4

[40] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 2

[41] O. Koller, H. Ney, and R. Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *CVPR*, 2016. 2

[42] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. *ICCV*, 2011. 2

[43] I. Laptev and T. Lindeberg. Space-time interest points. *ICCV*, 2003. 2

[44] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, 2008. 2

[45] A. Li, A. Jabri, A. Joulin, and L. van der Maaten. Learning visual n-grams from web data. In *ICCV*, 2017. 2

[46] Z. Li and D. Hoiem. Learning without forgetting. *PAMI*, 2017. 2

[47] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *ECCV*, 2018. 1, 2, 3, 4, 5, 6

[48] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2

[49] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 3

[50] J. Y.-H. Ng, J. Choi, J. Neumann, and L. S. Davis. Actionflownet: Learning motion representation for action recognition. In *WACV*, 2018. 1

[51] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 2

[52] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2

[53] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 2

[54] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014. 2

[55] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017. 2

[56] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 2

[57] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 2

[58] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 2

[59] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *PAMI*, 2011. 2

[60] Z. Shi, P. Siva, and T. Xiang. Transfer learning by ranking for weakly supervised object annotation. *arXiv preprint arXiv:1705.00873*, 2017. 2

[61] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012. 2

[62] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014. 2

[63] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *ICCV*, 2017. 1, 2, 3, 4

[64] VerbNet. VerbNet : A Computational Lexical Resource for Verbs. [Online] Available `https://verbs.colorado.edu/verbnet/`. 3

[65] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *arXiv:1705.06950*, 2017. 2, 3, 4, 8

[66] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. *CVPR*, 2018. 2, 8

[67] X. Wang and A. Gupta. Videos as space-time region graphs. *arXiv preprint arXiv:1806.01810*, 2018. 8

[68] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *PAMI*, 2008. 2

[69] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 2017. 8

[70] Y. Bian, C. Gan, X. Liu, F. Li, X. Long, Y. Li, H. Qi, J. Zhou, S. Wen, and Y. Lin. Revisiting the effectiveness of off-the-shelf temporal modeling approaches for large-scale video classification. *arXiv:1708.03805, 2017*, 2017. 3, 8

[71] M. Zolfaghari, K. S. Singh, and T. Brox. ECO: efficient convolutional network for online video understanding. *arXiv preprint arXiv:1804.09066*, 2018. 2, 8