A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning

Christoph Feichtenhofer  Haoqi Fan  Bo Xiong  Ross Girshick  Kaiming He
Facebook AI Research (FAIR)

Abstract

We present a large-scale study on unsupervised spatiotemporal representation learning from videos. With a unified perspective on four recent image-based frameworks, we study a simple objective that can easily generalize all these methods to space-time. Our objective encourages temporally-persistent features in the same video, and in spite of its simplicity, it works surprisingly well across: (i) different unsupervised frameworks, (ii) pre-training datasets, (iii) downstream datasets, and (iv) backbone architectures. We draw a series of intriguing observations from this study, e.g., we discover that encouraging long-spanned persistency can be effective even if the timespan is 60 seconds. In addition to state-of-the-art results in multiple benchmarks, we report a few promising cases in which unsupervised pre-training can outperform its supervised counterpart. Code is made available at https://github.com/facebookresearch/SlowFast.

1. Introduction

A series of recent methods on unsupervised representation learning from images [36, 12, 32, 9] are based on maximizing a similarity objective for different views of the same image under data augmentations [18, 89]. In addition to the artificial augmentations on images, videos can provide natural augmentations of visual content under various changing factors, such as motion, deformation, occlusion, and illumination. This work aims to generalize these image-based methods [36, 12, 32, 9] into space-time.

We study a simple objective that can be easily incorporated into these image-based methods. Our hypothesis is that the visual content is often temporally-persistent along a timespan in the video. This persistency may involve an action (e.g., a person dancing), an object (e.g., an individual person, who transitions from running to walking), and a scene (e.g., a room with people moving), covering short to long spans, with different levels of visual invariance (action, object, scene). Our objective simply encourages the visual representations in different clips of the same video to be similar. We empirically find that this objective works well across different unsupervised frameworks (MoCo [36], SimCLR [12], BYOL [32], SwAV [9]), either with or without using dissimilar (negative) samples.

Our objective is a natural generalization of crops in images [18, 89] to clips in videos. This allows us to make use of the recent unsupervised learning frameworks with minimal modifications. We aim to learn a high-level representation of the categorical semantics present in a video by enforcing persistency of the representation over space-time. We investigate factors such as the effective timespan, \( t \), between positives, and number of temporal clips, \( \rho \), to find that longer timespans (up to a minute) and multiple samples are beneficial for downstream performance (Fig. 1).

Our unsupervised training is performed on large-scale data, including Kinetics [47] (240k videos) and three versions of million-scale Instagram sets. In addition to standard linear probing, we evaluate representation quality on multiple classification and detection downstream datasets, e.g., Charades [75], Something-Something [31], and AVA [33].
Our results suggest that unsupervised pre-training can achieve competitive performance in videos, and it can surpass the supervised pre-training counterparts in a few cases. Finally, our study also reveals room for improvement along multiple directions.

In summary, our large-scale study involves the following five aspects:

(i) Four unsupervised learning frameworks (MoCo [36], SimCLR [12], BYOL [32], SwAV [9]) viewed from a unified perspective, and incorporated with a simple temporal persistency objective;

(ii) Three pre-training datasets, including the relatively well-controlled Kinetics [47] and the relatively “in-the-wild” Instagram sets at million-scale;

(iii) Six downstream datasets/tasks for evaluating representation quality;

(iv) Ablation experiments on different factors, such as temporal samples, contrastive objective, momentum encoders, training duration, backbones, data augmentation, curated vs uncurated, trimmed vs untrimmed, etc.; and

(v) State-of-the-art results of unsupervised video representation learning on established benchmarks, UCF-101 [77], HMDB51 [50] and Kinetics-400 [47].

2. Related Work

Unsupervised learning in images has been actively researched recently with approaches focusing on various pretext tasks related to color- or patch-based processing [67, 94, 17, 64], instance discrimination with contrastive objectives [18, 89, 83, 40, 41, 46, 36, 95, 12, 81] and ones that focus on positive pairs [8, 9, 32].

Unsupervised learning in videos has followed a similar trajectory with earlier methods focusing on predictive tasks based on motion, color and spatiotemporal ordering [29, 43, 1, 44, 78, 85, 60, 84, 58, 57, 21, 51, 86, 66, 22, 48, 91, 16, 87, 70, 45] and contrastive objectives with visual [74, 79, 34, 53, 28, 92] and audio-visual input [65, 4, 5, 49, 3, 68, 69].

Several recent ones [28, 35, 3, 68, 2, 71, 92, 62] relate to image-based approaches [36, 8, 12, 89]. With some of them using additional modalities of optical-flow [81, 35], audio [3, 68, 8, 2, 62] and text [79, 2] to transfer supervision from one modality to another.

In relation to these previous efforts, our work studies purely visual unsupervised learning from video and tries to compare the meta-methodologies on common ground.

Evaluation protocols and backbones in most image-based approaches have converged to ResNet-50 [39] encoders with ImageNet linear-classification protocol, and several smaller downstream tasks [36, 12, 32, 9] for evaluation. In video understanding research, the field has not yet converged and is using different backbones with focus on fine-tuning performance on two relatively small datasets [77, 50]. We investigate this aspect by looking at different encoders and 6 different downstream benchmarks for evaluation.

3. Approach

The objective of this work is to study several recent unsupervised representation learning methodologies to train a spatiotemporal encoder \( f_\theta \), exploring implementation details and comparing them on a common ground to measure their efficacy in video understanding. We focus on two contrastive approaches using positive and negative samples: SimCLR [12] and MoCo [36], as well as two approaches that solely rely on positives, BYOL [32] and SwAV [9] (Sec. 3.2).

These approaches were originally presented for learning image representations, and they all share the objective of learning invariant features across different views (crops/augmentations) of the spatial image input. In this paper, this idea is extended to the temporal domain. Our core idea is to learn an encoder \( f_\theta \) that produces embeddings which are persistent in space-time, over multiple \( \rho \) temporally distant clips of the same video. This is related to Slow Feature Analysis [88] where the objective is to minimize the representations’ temporal derivative over the input. The general idea of learning temporally persistent features is not new and has been proposed in the past with similar motivation e.g., [6, 61, 29].

3.1. Persistent temporal feature learning

Our framework takes different augmented clips \( x \) of an unlabeled video and passes them through an encoder \( f_\theta \) with weights \( \theta \) to obtain corresponding embeddings \( q = f_\theta(x) \). The encoder is spatiotemporal ConvNet, by default a ResNet-50 (R-50) [39], Slow-only pathway of SlowFast Networks [20], which is a 3D ResNet-50 [39] without temporal pooling in convolutional feature maps, followed by an MLP projection head, that produces and output of dimension \( d \).

The input clips are stacks of RGB frames of size \( 3 \times T \times S \) for temporal \( \times \) spatial dimensions, which are sampled with temporal stride \( \tau \), i.e., the encoder processes only one out of \( \tau \) frames of the raw video. Therefore, \( T \times \tau \) define the timespan and resolution of the encoder.

Given a minibatch of \( B \) videos, our framework creates a set of \( \rho B \) positive examples by sampling \( \rho \) clips from the videos. The learning methodologies studied in this section maximize similarity of a “query” sample \( q \) with a set of positive “key” samples \( \{k^+\} \) that are encoded versions of different clips of the same video as \( q \) is computed from. Fig. 1 illustrates an example where \( \rho=3 \) clips are used.

The next section describes how the contrastive and non-contrastive unsupervised representation learning methodologies are exemplified.
3.2. Unsupervised learning frameworks

Contrastive learning maximizes the similarity of a sample $q$ with positive ones $\{k^+\}$ and minimizes similarity to negative ones $\{k^-\}$. The contrastive approaches in this paper use the InfoNCE [83] objective,

$$L_q = -\log \frac{\sum_{k \in \{k^+\}} \exp(\text{sim}(q, k)/\alpha)}{\sum_{k \in \{k^+, k^-\}} \exp(\text{sim}(q, k)/\alpha)},$$

with $\alpha$ being a temperature hyper-parameter for scaling and $\{k^+\}$ are embedded clips of the same video as $q$. All the embeddings are $\ell_2$ normalized and dot product (cosine) similarity is used to compare them $\text{sim}(q, k) = q^T k / \|q\| \|k\|$. 

**SimCLR** [12] (Fig. 2a) uses the embeddings of clips of other videos in the minibatch as negatives $\{k^-\}$.

**MoCo** [36] (Fig. 2b) is a method that uses an explicit momentum encoder which parameters, $\theta_m$, are moving average $\theta_m \leftarrow m \theta + (1 - m) \theta$ with $m$ a momentum parameter. In eq. (1) MoCo uses this encoder to compute the positive embeddings $\{k^+\}$ from clips of the same video as $q$, and negative embeddings $\{k^-\}$ are taken from a queue that stores embeddings of clips from previous iterations. There is no backpropagation into the momentum-encoder weights $\theta_m$.

**BYOL** [32] (Fig. 2c) can be viewed as a form of MoCo that does not use negative samples, but an extra predictor MLP with weights $\theta_p$, which is stacked on top of $f_0$’s MLP head. For a sample $q = f_\theta_p(f_0(x))$, BYOL minimizes negative cosine similarity,

$$L_q = -\sum_{k \in \{k^+\}} \text{sim}(q, k) = -\sum_{k \in \{k^+\}} q^T k^+ / \|q\| \|k^+\|,$$

with $\{k^+ = f_\theta_m(x^+)\}$ being embedded clips $x^+$ from the same video as $q$, encoded with momentum weights $\theta_m$.

**SwAV** [9] (Fig. 2d) can be viewed as a form of SimCLR that does not use negative samples. SwAV first performs a linear mapping of the positive embeddings $q, k^+$ to learned prototypes $\tilde{q}, \tilde{k}^+$ and then transforms the targets with an extra Sinkhorn-Knopp ($SK$) step. Then the SwAV loss is

$$L_q = D_{KL}(\tilde{q} \parallel SK(\tilde{k}^+)),$$

where $D_{KL}$ is the Kullback-Leibler divergence and gradients are not back-propagated through the $SK$ operation.

Compared to SimCLR and MoCo, in BYOL and SwAV, $q$ and $k$ are not typical “query” and “key” samples (but rather “source” and “target” samples); however, for consistency we use $q, k$ terminology in notation for all methods.

**Implementation specifics.** We implement the methods with a symmetric loss, as in original SimCLR, BYOL and SwAV, where every input clip is used to produce a loss (and gradient) signal. For each of the $\rho \geq 2$ clips, we compute $q$, while all other $\rho-1$ clips of the same video are used as $\{k^+\}$ to evaluate sub-loss $L_q$ and the symmetric loss is the average over all $\rho$ sub-losses. Thus, for MoCo and BYOL, every input clip is processed by both encoders.

For MoCo and BYOL, our symmetric loss is aggregated sequentially which implies that memory consumption for $\rho > 2$ equals to a single clips’ forward and backward pass, since these methods do not backpropagate through the momentum encoder. For SimCLR and SwAV the overall loss is evaluated in parallel across all clips and therefore memory consumption grows linearly with the number of clips used.

All details on implementation and pre-training are in §B.1.

4. Experiments

**Datasets.** Unless otherwise noted, we perform unsupervised pre-training on Kinetics-400 [47] (K400) with ~240k training videos in 400 human action categories.
To study learning from “in-the-wild” videos from the web, we pre-train the methods on Instagram videos: IG-Curated [24], a dataset with hashtags similar to K400 classes; IG-Uncurated which has videos taken randomly from Instagram; and IG-Uncurated-Short which is similar, but has constrained duration. Each dataset has 1M videos.

Table 1 shows dataset statistics of all datasets used for unsupervised pre-training. Most of Kinetics videos are of 10 seconds in duration. IG-Curated is a dataset with Instagram videos that have an average duration of 26.3 seconds and a standard deviation of 29.8 seconds. The maximum duration is 60s. IG-Uncurated contains videos taken randomly from Instagram, with larger deviation in length and maximum duration of 10 minutes (600s). IG-Uncurated-Short is a dataset consisting of random Instagram videos that have a duration between 10 and 16 seconds, to study the effect of a fixed duration and the assumption that short videos may hold more useful information for pre-training.

Evaluation protocols. For evaluation we use two protocols.

The first one is common to evaluate unsupervised image representations [36, 12]. It validates the linear classifier performance based on frozen encoder features that are taken from the global average pooling layer. We report top-1 classification accuracy (%) on the K400 validation set.

The second protocol reports fine-tuning accuracy on the first split of the UCF101 dataset [77] which contains 13k videos in 101 human action classes; this is a common procedure used to evaluate unsupervised video representations. Finally, we also report fine-tuning accuracy on AVA [33], Cha redes [75], Something-Something [31] and HMDB51 [50].

Architecture. By default, we use a R-50 [39] following the Slow pathway in [20] with clips of $T=8$ frames sampled with stride $\tau=8$ from 64 raw-frames of video. The supervised performance for training 200, 400, 800 epochs on K400 is 74.7%, 74.3% and 72.7%, respectively, and does not improve for training longer due to overfitting.

Implementation details. We follow default settings in video classification [20]. Specifics on the approaches, their training and evaluation and the impact of implementation on performance are provided in §8 and §A.3, respectively.

4.1. Persistent temporal learning

Here, we investigate the impact of learning spatiotemporal vs. only spatial persistent features. Table 2 shows the accuracy of the four methods when trained for 200 epochs on K400, and evaluated on K400 (finedtuned), i.e. our default setting.
4.2. Timespan between positives

All experiments with $\rho \geq 2$ so far were using global temporal sampling of positives, which means that the clips can be sampled at unconstrained temporal locations from the input video. This might be counter-productive because if there is a long duration that has passed between a pair of positive clips they might no longer share the same semantic context for learning high-level features corresponding in time.

| $t_{\max}$ in seconds | 0 | 2 | 3 | 4 | 5 | 8 | 10 |
|------------------------|---|---|---|---|---|---|----|
| K400 acc in %          | 60.6 | 65.2 | 65.7 | 65.8 | 65.8 | 65.6 | 65.8 |
| (a) Dataset: K400, 200 epochs training. |
| $t_{\max}$ in seconds | 4 | 8 | 16 | 32 | 60 |
| K400 acc in %          | 62.7 | 63.1 | 63.1 | 63.9 | 64.1 |
| (b) Dataset: IG-Curated-1M, 50 epochs training. |
| $t_{\max}$ in seconds | 12s | 24 | 36 | 48 | 600 |
| K400 acc in %          | 59.3 | 59.2 | 59.9 | 59.6 | 58.9 |
| (c) Dataset: IG-Uncurated-1M, 50 epochs training. |

Table 4. Maximum frame distance for positives. Method: BYOL, $\rho = 2$. Training is surprisingly robust with increasing accuracy for increased distance between samples. Accuracy only (mildly) degrades when sampling positives that are more than 36 seconds apart when using uncurated (random) videos.

This experiment is concerned with the maximum distance between the positive training samples. We use BYOL pre-training on K400, IG-Curated-1M and IG-Uncurated-1M and report 400 linear readout accuracy in Table 4.

Table 4a shows performance for increasing the maximum temporal distance between positives in K400 pre-training. It can be seen that using positives from the same time ($t_{\max}=0$) degrades performance by less than that other than performance is relatively robust up to global sampling of positive clips from the whole video ($t_{\max}=10s$). This is interesting as it seems that a long-temporal correspondence objective does not hurt performance (but also does not boost it).

Table 4b shows performance for increasing the temporal distance between positive samples on IG-Curated-1M. This dataset has a maximum duration of 60 seconds; statistics are in Table 1. Table 4b reveals that increasing the maximum duration between positive pairs is beneficial for performance and unrestricted sampling of positives is the best with 64.1% top-1 accuracy for evaluation on K400. This is especially interesting, as it shows that even longer videos benefit from global sampling. There is no benefit from restricting the time window of positives, which can be interpreted as the objective of learning extremely-slow features [88] that do not change over 60 seconds of video. Long-temporal-distance samples might also increase robustness of the model by providing “hard-positive” samples for learning. Note that here the videos are still sampled according to hashtags related to K400 classes [24]; therefore, the conjecture might be biased.

Finally, we are looking at the IG-Uncurated-1M dataset which consists of a random sampling of 1M videos from Instagram. These videos can be between 0.5s and 10 minutes of duration. Most of the videos however are much shorter than 10 minutes, with a mean duration of 35.3 seconds and a standard deviation of 38.4 seconds (Table 1). For this data, Table 4c shows the results of progressively increasing the maximum timespan between positive samples. It can be observed that increasing the maximum distance between positives up to 36 seconds is beneficial and beyond that performance decreases, but only slightly, even when performing global sampling of positives (the default).

4.3. Backbone architectures

So far all experiments were using a R-50, 8×8 Slow pathway [39, 20] as backbone. The next set of ablations studies different architectures for the spatiotemporal encoder.

Table 5 compares different backbones for usage with MoCo in our default setting ($\rho=2$, 200 epoch pre-training on K400). From left to right, the table shows the input duration $T$, sampling-rate $\tau$, FLOPs (at 224×224 spatial resolution) and parameters of these backbones, as well as the average duration for training one iteration of the MoCo algorithm (measured on a single machine with 8 V100 GPUs in PySlowFast [19] and torchvision decoder), the supervised performance on K400 and UCF101 (finetuned from K400), as well as the downstream performance for K400 linear evaluation and UCF101 finetuning.

The first observation in Table 5 is that for the Slow architecture [20], using shallower (R-18) or deeper (R-101) networks can influence supervised and downstream performance in a sizable manner, with MoCo, K400 evaluation benefiting from more parameters. Doubling the input frame-rate (8×8 → 16×4) boosts accuracy on UCF101.

The second observation is that R2+1D [82] has a large gap on Kinetics (71.7% supervised vs. 57.2% unsupervised), while being remarkably strong on UCF101 (93.7%). This gap is also observed for S3D-G [90]. The reason for this might be that UCF101 is a small dataset which is easy to overfit and can benefit from fewer parameters.
4.4. Uncurated data and video duration

In Table 6 we show the performance of all four methodologies on IG-Curated-1M (a), IG-Uncurated-1M (b) and IG-Uncurated-Short-1M (c) for pre-training with 50 and 200 epochs. We make the following observations:

(i) Among the methods MoCo performs the best with e.g. 69.0% vs. second-best 64.3% of SwA V on curated data (a).

(ii) MoCo and SwA V scale the best for training longer, gaining roughly 3-4% for 200ep vs. 50ep.

(iii) On uncurated data, MoCo and SwA V perform ~1% better on the unconstrained duration videos in Table 6b.

(iv) BYOL and SimCLR show better performance on IG-Uncurated-Short (10-16s videos) in Table 6c, seemingly benefiting from shorter videos, but there is no clear benefit from either longer or shorter duration among all methods.

(v) BYOL degrades performance for training longer which might be due to the requirement of different hyper-parameters for different schedules (as noted in Sec. 4.1).

We will return to this point in §A.1, where we show that increasing clips-size $\rho$ can overcome this issue in BYOL, along with further studies on the trade-off against training more epochs, and dataset scale.

4.5. Data augmentations

Importance of augmentations. Augmentations can have a major impact on visual unsupervised feature learning [12, 14]. In Fig. 3, we ablate spatial cropping (S), temporal clipping (T) and radiometric color (C) augmentations from the four unsupervised learning methods (e.g. “T S C” are the baselines using all augmentations and removing “S C” equals $\rho=1$ in Table 2). We make three main observations:

(i) Among the methods, MoCo and BYOL perform most robust for using fewer augmentations; their advantage over SimCLR and SwA V might be related to the momentum encoder which can provide extra augmentation in training.

(ii) When minimizing the augmentations by resizing the shorter size of the video to the input size of 224 and only cropping along the long side of the video (Base in Fig. 3), MoCo still provides 42.2% K400 linear accuracy, over BYOLs’ 32.4%, showing an advantage of the contrastive loss in a weak augmentation scenario.

(iii) Among the augmentations, learning temporal (T) persistence, has the largest impact on performance, except for MoCo which benefits more from color (C) (incl. grayscale) augmentations. Especially SimCLR and SwAV show significant drops in performance when removing T, i.e. when extracting positive clips from the same instance in time.

In the remainder of this section, we explore using stronger augmentations than the default ones in previous experiments. We perform the ablations with MoCo in the basic setting of $\rho = 2$, 200 epochs K400 pre-training.

| color strength | grayscale probability | temporal difference | fps jitter | accuracy K400 | accuracy UCF101 |
|----------------|-----------------------|---------------------|-----------|--------------|----------------|
| 0.5            | 0.2                   | ✓                   | 65.8      | 91.0         |
| 0.75           | 0.2                   | ✓                   | 66.0      | 92.1         |
| 1.0            | 0.2                   | ✓                   | 65.8      | 91.2         |
| 0.5            | 0.4                   | ✓                   | 65.5      | 91.0         |

Table 7. Radiometric augmentation. Method: MoCo. 200 epochs, $\rho = 2$. Dataset: K400. Stronger color augmentation in K400 pre-training can especially benefit UCF101 (+1.3%).

Stronger color augmentation. In Table 7 color strength of 0.5 indicates the default one for MoCo [14], 0.75 and 1.0 increase the strength of randomly jittering brightness, contrast, saturation and hue proportionally.

Table 7 shows that increasing it to 0.75 can improve K400/UCF101 accuracy. Increasing the random grayscale probability from 0.2 to 0.4 does not provide an improvement on either of the datasets. However, using a temporal-difference augmentation which randomly (with probability 0.2) first converts the frames to grayscale and then subtracts them across time, can increase K400 accuracy by 0.4%. Finally, using frame-rate jittering of ±50% of the original frame-rate does not improve K400 but UCF101 slightly.
4.6. Alternative downstream tasks

The gap between K400 and UCF101 accuracy in Sec. 4.3 question if solely looking at typical evaluation of UCF101 (or the smaller HMDB51) is enough to identify and rank approaches for unsupervised learning in video.

Table 10 studies several new downstream tasks for unsupervised representation learning in video. We use our MoCo, SimCLR, BYOL and SwA V models trained with \( \rho = 3 \) for 200 epochs on K400 and evaluate their performance by finetuning on Charades [75], AVA [33], or Something- Something [31] (in addition to the K400 linear readout performance and UCF101 performance reported in Table 2). Details on implementation are given in §B.

The first two rows in Table 10 show the two main competitors for this evaluation: (i) training from scratch on the datasets and (ii) K400 pre-training. First, we observe that the supervised pre-trained backbones outperform the train-from-scratch counterpart significantly, as expected.

**Downstream datasets.** For K400 pre-training and linear evaluation, its supervised counterpart has an advantage between 12.7% and 6.4% top-1 accuracy among the methods.

On UCF101 unsupervised pre-training is only 1% lower than the supervised counterpart for BYOL (the strongest).

On AVA short-term action detection we observe that the BYOL pre-trained model is able to outperform the supervised counterpart by +1.2% mAP, when using the same, fixed region proposals [20]. This result is significant, as e.g. switching from K400 to K600 (nearly double the size of K400) pre-training on AVA leads to a smaller gains in performance [20]. Overall this is a surprising result as the tasks in K400 and AVA are similar [52], only that the temporal granularity of the actions in AVA is finer while their semantic granularity is coarser; e.g. “shoot” in AVA vs. “playing paintball” in Kinetics, which might be better captured by the BYOL objective which solely works on positive temporal samples of a video, without contrasting them to other videos (“shoot” might be a positive appearing in many different videos and contrasting them could be harmful to downstream performance). This line of thinking is supported with MoCo’s (contrastive objective) performance that is 3.1% worse than BYOL on AVA. Similarly, SimCLR (contrastive) is worse than SwA V (non-contrastive) when benchmarked on AVA.

On Charades, long-term action classification, we observe the opposite. Here, the contrastive MoCo is clearly the best performer with 33.5% mAP (close to the supervised pre-training performance of 34.7% mAP), while the non-contrastive BYOL is 12.5% lower. Similarly, now SimCLR (contrastive) is better than SwAV (non-contrastive). Compared to AVA, Charades is a temporally less localized dataset containing activities that need to be recognized from a longer temporal range video, for which contrastive pre-training appears to be outperforming the non-contrastive variants.
Table 10. Downstream benchmarks: We use linear evaluation on K400 and finetuning accuracy on the other datasets. 200 epochs. $\rho=3$.

| method     | pre-train | linear protocol | UCF101 | AVA (mAP) | Charades (mAP) | SSv2 |
|------------|-----------|-----------------|--------|-----------|----------------|------|
| supervised | scratch   | K400            | 74.7   | 68.8      | 11.7           | 7.4  |
| supervised | K400-240K | -               | 94.8   | 22.2      | 34.7           | 52.8 |
| SimCLR     | K400-240K | 62.0 (−12.7)    | 87.9 (−6.9) | 17.6 (−4.6) | 11.4 (−23.3) | 52.0 (−0.8) |
| SwAV       | K400-240K | 62.7 (−11.5)    | 89.4 (−5.4) | 18.2 (−4.0) | 10.7 (−24.0) | 51.7 (−1.1) |
| BYOL       | K400-240K | 68.3 (−6.4)     | 93.8 (−1.0) | 23.4 (+1.2) | 21.0 (−13.7) | 55.8 (+3.0) |
| MoCo       | IG-Curated-1M | 67.3 (−7.4) | 92.8 (−2.0) | 20.3 (−1.9) | 33.5 (−1.2) | 54.4 (+1.8) |
| MoCo       | IG-Uncurated-1M | 69.9 (−4.8) | 94.4 (−0.4) | 20.4 (−1.8) | 34.9 (+0.2) | 54.5 (+1.8) |
| MoCo       | IG-Curated-1M | 66.0 (−8.7) | 92.9 (−2.1) | 20.5 (−1.7) | 31.3 (−3.4) | 53.2 (+0.4) |

On Something-Something v2 (SSv2 in Table 10), all the methods perform strong, with BYOL pre-training showing the largest gain of +3% over supervised pre-training on Kinetics (55.8% vs. 52.8% top-1 accuracy).

Pre-training sets: Kinetics vs. IG. Next, we experiment with pre-training on videos from the web. We first investigate IG-Curated-1M [24], which is a dataset that has been collected with hashtags that are similar to Kinetics labels. This data is a 1M subset of the original 65M introduced in [24]. Using this data (penultimate row in Table 10) can excel the performance of MoCo with K400 pre-training, which has a training set of 240K samples (roughly 4.2× smaller), and surprisingly even outperforms pre-training on K400 linear readout itself (69.9% vs. 67.3% accuracy).

Second, we ablate the effect of using uncached videos, with IG-Uncurated-1M which are purely random videos taken from the web. On most downstream tasks performance shown in the last row of Table 10 is equal or only slightly lower than pre-training on K400. Specifically, MoCo changes by -1.3% on K400 (as expected), +0.1% on UCF, +0.2% on AVA, -2.2% on Charades and -1.2% on Something-Something v2. This is an encouraging result for unsupervised learning, as only ~4.2× the number of videos but random ones are required to match the performance of supervised K400 pre-training on the UCF101 and AVA.

Overall, our results indicate that unsupervised pre-training can be a new paradigm for all of these downstream tasks, for which supervised pre-training is the de-facto standard to achieve best performance. Further, the large difference in performance for pre-training methodologies and objectives (e.g. contrastive/non-contrastive) revealed in the light of these benchmarks signals large room for future work.

4.7. Comparison to previous work

In a final experiment we take the best model from Table 9 and compare it with the state-of-the-art using the commonly used protocols on UCF101 and HMDB51 (across all 3 train/val splits) and K400. In Table 11 we show the results.

The strongest previous approaches are using multi-modal input, Vision “V”, Audio “A”, Text “T”, to train a contrastive objective across modalities; XDC [3] performs DeepCluster [8] on (V+A), CVRL [71], GDT [68] and MMV [2] use an objective similar to SimCLR on (V), (V+A), and (V+A+T), with the latter training on a Audioset (AS) [23] and HowTo100M (HT) [59], and CoCLR [35] can be seen as a variant of MoCo on rgb and optical-flow input.

In comparisons, our best performing model \(\rho_{BYOL}\), which is BYOL trained with temporal persistency over \(\rho=4\) clips, (cf. Tables 2 & 9), provides a substantial performance gain over the best published method [35]: +5.7% and +12.1% top-1 accuracy on UCF101 and HMDB51 (using identical backbone and pre-training data).

On K400 linear evaluation with the same data and R-50, Slow pathway [20] as backbone, our approach outperforms the concurrent CVRL [71] by +5.4% accuracy.

5. Conclusion

This paper has studied four meta-methodologies for unsupervised learning from video. Our findings include that it is beneficial to sample positives with longer timespans between them, contrastive objectives are less influential than momentum encoders, and training duration, backbones, video augmentation and curation are all critical for good performance. Our resulting models which learn persistent features across augmented spacetime clips set a new state-of-the-art.

We observed that linear readout on Kinetics is a good indicator of the performance on other datasets and that unsupervised pre-training can compete with the supervised counterpart on several datasets, but there is room for improvement. We hope that our baselines will foster research and provide common ground for future work.
This appendix provides additional material: §A contains further results on “in-the-wild” data (§A.1), Kinetics-600 (K600) [10] and Kinetics-700 (K700) [11] data (§A.2) and on the effect of key implementation details (§A.3).

§B contains additional implementation details for: Unsupervised pre-training (§B.1), and downstream evaluation in Kinetics (§B.2), AVA (§B.3), Charades (§B.4), Something-Something V2 (§B.5), UCF101 (§B.6), HMDB51 (§B.7).

A. Additional Results

A.1. Scaling “in-the-wild” data

As a follow-up experiment Table 12 compares training BYOL longer (200ep) to increasing its clips-size \( \rho \) but not training longer (50ep). For both (a) curated and (b) random data, this results in a significant gain of performance.

| \( \rho \) | ep | BYOL K400 | UCF101 | BYOL K400 | UCF101 |
|---|---|---|---|---|---|
| 2 | 50 | 64.1 | 93.5 | 2 | 50 | 58.9 | 90.1 |
| 2 | 200 | 60.2 | 92.7 | 2 | 200 | 57.9 | 91.6 |
| 4 | 50 | 67.7 | 94.5 | 4 | 50 | 63.8 | 91.8 |

(a) IG-Curated-1M. (b) IG-Uncurated-1M.

Table 12. More epochs (ep) vs. more clips (\( \rho \)). Longer training degrades performance for BYOL, but increasing \( \rho \) does not.

We also explore an experiment for increasing the clip-size in MoCo and training longer (as MoCo works stable for more epochs). Table 13 shows the results. It can be observed that increasing the number of clips from \( \rho = 2 \) to \( \rho = 3 \) can increase the results by 1.6%/0.9% K400 and 0.4%/1% on UCF101 for 100/200ep training. Going to \( \rho = 4 \) brings further gain. In terms of efficiency, increasing \( \rho \) is both more accurate and faster than increasing the number of epochs, e.g. training MoCo (\( \rho = 3 \), 100ep) takes only 63% of the duration that MoCo (\( \rho = 2 \), 200ep) requires.

| ep | MoCo (\( \rho = 2 \)) K400 | UCF101 | MoCo (\( \rho = 3 \)) K400 | UCF101 | MoCo (\( \rho = 4 \)) K400 | UCF101 |
|---|---|---|---|---|---|---|
| 100 | 67.5 | 93.3 | 69.1 | 93.7 | 69.8 | 94.9 |
| 200 | 69.0 | 93.4 | 69.9 | 94.4 | 69.9 | 94.9 |

Table 13. More epochs (ep) vs. more clips (\( \rho \)). Dataset: IG-Curated-1M, \( \rho = 2 \). Training longer is less effective than increasing the number of temporal clips per iteration (\( \rho \)).

Finally, we remark that the IG-Curated-1M is subsampled such that the hastags are uniformly distributed (roughly balanced). Therefore this dataset is matching K400 in terms of content and distribution. We revisit this point next by investigating the effect of scale, curation and balancing of the video data.

In this experiment, we increase the scale of the data from 128K to 1M distinct videos. We increase dataset size (number of videos) for IG-Curated [24], IG-Curated-Unbalanced [24] (which has random class distribution), and IG-Uncurated (which are random IG videos). The experiment with 200-epoch MoCo with \( \rho = 2 \), linear protocol downstream evaluation on K400 linear evaluation protocol. The dotted line represents a linear trend. Method: MoCo, 200 epochs, \( \rho = 2 \).

![Figure 4. Data scale and curation. We increase dataset size (number of videos) for IG-Curated, IG-Curated-Unbalanced, and IG-Uncurated. By using \( 4 \times \) the number of videos, IG-Uncurated approaches the heavily curated Kinetics (K400) pre-training on K400 linear evaluation protocol. The dotted line represents a linear trend. Method: MoCo, 200 epochs, \( \rho = 2 \).](image)

A.2. Scaling Kinetics data

As referenced in Sec. 4 of the main paper, Table 14 shows a series of extra results for pre-training on the larger-scale Kinetics-600 (K600) [10] and Kinetics-700 (K700) [11] datasets, and is analyzed next: The first row of the table shows supervised training on the respective datasets, where UCF101 has two entries, one for training-from-scratch and one for using K400 as pre-training.

For the experiments we focus on our temporally persistent MoCo algorithm and, as in the main paper, evaluate Kinetics with the linear classification protocol and UCF101 by finetuning all weights. The first unsupervised row in Table 14 shows our best K400 pre-trained MoCo (\( \rho = 4 \)) model, achieving 69.0%, 70.0%, 54.2% and 93.6% on K400, K600, K700 and UCF101, respectively (this is the model with strong augmentations from Table 10 of the main paper).
Table 14. **Dataset scale**: Configuration: backbone: R-50, Slow $8 \times 8$, 200 epochs. Our approach, MoCo ($\rho=4$), is able to approach supervised pre-training on the popular UCF101 evaluation protocol, but there remains a gap for the linear protocol on K400, K600 and K700.

| method     | pre-train data | #videos | K400 | K600 | K700 | finetune UCF101 |
|------------|----------------|---------|------|------|------|-----------------|
| supervised | scratch        |         | 74.7 | 78.1 | 65.2 | 68.8            |
| MoCo ($\rho=4$) | K400 240k |         | 69.0 | 70.0 | 54.2 | 93.6            |
| MoCo ($\rho=2$) | K600 387k |         | 69.6 | 70.7 | 55.1 | 92.7            |
| MoCo ($\rho=4$) | K600 387k |         | 71.5 | 72.8 | 57.7 | 94.5            |
| MoCo ($\rho=2$) | K700 522k |         | 70.0 | 71.4 | 56.2 | 92.8            |
| MoCo ($\rho=4$) | K700 522k |         | 71.7 | 73.2 | 58.1 | 94.8            |

The next row shows MoCo trained on K600 with a temporal persistency objective across two clips, $\rho=2$. This version is able to slightly outperform the K400 pre-trained variant on all datasets, except UCF101. Directly comparing this version with learning temporal persistency across $\rho=4$ clips can significantly increase accuracy on all datasets by $\sim$2%.

The final two rows of Table 14, show the same two models when pre-trained on K700. Here, we see that going from K400 to K700 increases accuracy by 2.7%, 3.2% and 3.9%, 1.2% on K400, K600, K700 and UCF101, respectively.

Overall the experiments suggest clear benefits of using larger-scale datasets for unsupervised pre-training and room for improvement under the linear classification protocol, especially when evaluated on larger datasets.

### A.3. Key implementation specifics

While the full implementation details of all four meta-methodologies are provided in §B.1, we want to discuss the most impactful ones, which we found critical to achieve good performance in their realizations, throughout this section.

| $m_{\text{base}}$ | n/a | 0.988 | 0.990 | 0.992 | 0.994 | 0.996 |
|-------------------|-----|-------|-------|-------|-------|-------|
| acc.              | 64.5| 65.5  | 65.5  | 65.6  | 65.8  | 65.1  |

**Table 15. Momentum annealing for MoCo.** Dataset: K400, 200 epochs, $\rho=2$. Using cosine-annealing of the momentum brings gains of $\sim$1% accuracy. We use 0.994 as default for MoCo.

**Momentum annealing.** BYOL is using an annealing of the rate at which parameters of the momentum encoder $\theta_m$, that are a moving average, with momentum $m$, of the trained encoder $\theta$. During training BYOL starts with a momentum of $m_{\text{base}}=0.996$ and increases it to 1 with a cosine annealing $m = 1 - (1 - m_{\text{base}}) \cdot \cos(\pi k/K) + 1)/2$ with $k$ the current iteration and $K$ the maximum number of training iterations [32] (this is unrelated to the learning rate decay).

By default MoCo, is using a fixed momentum of $m = 0.999$ during training. In Table 15, we ablate the positive (or negative) effect of using momentum annealing with different starting rates $m_{\text{base}}$ for MoCo. We observe that not using any annealing (n/a) produces 64.5% accuracy and using momentum annealing can boost this performance by $\sim$1%, while being relatively stable for different values of $m_{\text{base}}$. Consequently, we are using momentum annealing with $m_{\text{base}} = 0.994$ for all our MoCo experiments.

**Normalization and optimization.** Here, we present normalization specifics that we found critical to achieve good performance in the underlying implementation of the methods: SimCLR, BYOL and SwAV are using synchronized Batch Normalization (BN) [42] statistics (SyncBN) across 8 GPUs during training, batch-normalization after every MLP layer (MLP-BN), and a large-batch optimizer (LARS) [93]. LARS adaptively scales the learning rate for each individual parameter by using the ratio between gradient and parameter magnitudes. MoCo is not using these components (None) by default. In Fig. 5 we illustrate the results. It shows accuracy on K400 linear readout, if step-by-step adding these specifics to the methods. We make the following observations:

(i) Using None of the augmentations provides best performance for MoCo (its default) but significantly degrades BYOL, SimCLR and SwAV. Here, it is worth noting that BYOL provides decent accuracy of 32.9% without SyncBN, LARS and any BN in the MLP.

(ii) Adding LARS optimizer reduces performance in MoCo and BYOL, while having a boost of around 10% for both SimCLR and SwAV. It is interesting, that solely using a more advanced optimizer, which adapts the learning rates of the weights according to their gradient magnitudes, decreases performance in methods using a momentum encoder (MoCo, BYOL), but boosts it without (SimCLR, SwAV).

(iii) further adding SyncBN and MLP-BN increases BYOL performance dramatically; this related to recent studies [73] which suggest that normalization is important to achieve good performance using BYOL.
(iv) While BYOL, SimCLR and SwAV do show further gains for adding SyncBN and MLP-BN, MoCo shows no significant change for using SyncBN, and degrades drastically in performance for using BN in the MLP-head.

**Projection MLP.** It has been shown that using a deeper projection MLP in pre-training can increase the accuracy of the resulting representations for image classification [12, 14, 13]. Here, we investigate the effect of more hidden layers for video classification, across all four meta architectures. The results are shown in Table 16 and discussed next.

(i) MoCo achieves a significant gain of 1.2% on K400 for using a 3-layer (2 hidden layers) MLP vs. using a 2-layer MLP and there is no gain for using a 4th layer. UCF performance appears stable to this modification. The gain is in line with results in image classification [14].

(ii) For BYOL, which has an additional Predictor MLP, with weights $\theta_p$ (see Fig. 2c), we ablate two dimensions: increasing the projection depth, and the prediction depth. Our results show that using 3-layer projection vs. 2-layer does not affect performance on K400, and has a decay of -0.7% on UCF101. Increasing also the depth of the predictor from our default value of 2 to 3 layers will lead to a significant decrease of -2.2% and -2.5% on both K400 and UCF101.

(iii) SimCLR, shows similar behavior as MoCo: A consistent gain for using 3 projection layers (+1.5% on K400, +0.5% on UCF101), and no further gain for a 4-layer MLP.

(iv) SwAV shows continuing gains on K400 for adding more MLP layers, +1.3% for going from 2 to 3 and another +0.4% for 4-layer MLP; however, its UCF101 performance is decaying with more projection layers.

Overall, Table 16 suggests that K400 linear evaluation accuracy generally benefits from deeper projection heads, while the performance for fine-tuned UCF101 downstream performance is relatively unchanged and rather shows a decaying effect for deeper MLPs. When studying the training complexity for pre-training, which we measure as floating point operations (FLOPs) and Parameters for the full training architecture (encoders + MLPs), Table 16 shows that FLOPs are mostly unchanged by deeper MLPs (as they operate on feature maps of size $1 \times 1 \times 1$), but parameters increase leading to large models especially for momentum encoder based approaches (MoCo and BYOL).

| method | MLP layers | training FLOPs | Param | K400 | UCF101 |
|--------|-----------|---------------|-------|------|--------|
| MoCo   | 2         | 41.74G 72.2M  | 64.6  | 91.3 |
|        | 3         | 41.74G 80.6M  | 65.8  | 91.0 |
|        | 4         | 41.75G 89.9M  | 65.7  | 91.0 |
| BYOL   | 2, predictor: 2 | 41.75G 86.4M  | 65.8  | 92.7 |
|        | 3, predictor: 2 | 41.77G 119.9M | 65.8  | 92.0 |
|        | 3, predictor: 3 | 41.78G 153.5M | 63.6  | 90.2 |
| SimCLR | 2         | 41.74G 36.1M  | 59.0  | 88.4 |
|        | 3         | 41.75G 40.3M  | 60.5  | 88.9 |
|        | 4         | 41.75G 44.5M  | 60.6  | 88.5 |
| SwAV   | 2         | 41.74G 36.2M  | 60.3  | 88.1 |
|        | 3         | 41.75G 40.4M  | 61.6  | 87.3 |
|        | 4         | 41.75G 44.6M  | 62.0  | 87.1 |

Table 16. Varying depth of MLPs. Dataset: K400, 200 epochs, $\rho=2$. Training complexity is measured in floating point operations (FLOPs) and Parameters. Accuracy is reported as linear evaluation (K400) and fine-tuning (UCF101) of the backbone without MLPs.

**MoCo details.** We use a queue storing 65536 negatives and shuffling BN to avoid intra-batch communication among samples [36]. We use a 3-layer (2 hidden layers, ablation in Table 6 of the main paper) projection MLP with hidden dimension 2048, ReLU activation [63] and no BN. Other hyperparameters are as in [36, 12, 9, 32]. The temperature parameter $\tau$ is set for each method to $\tau_{\text{MoCo}} = 0.4$, and $\tau_{\text{SimCLR}} = \tau_{\text{BYOL}} = \tau_{\text{SwAV}} = 4.8$. We apply (LARS) [93] (except for bias and BN parameters [32]), with trust coefficient of 0.001, for BYOL, SimCLR, and SwAV training. The SGD weight decay is $10^{-4}$ for MoCo and $10^{-6}$ for BYOL, SimCLR and SwAV. The temperature parameter $\alpha = 0.1$ for MoCo, SimCLR and SwAV. The projection MLP output dimensions are $d_{\text{MoCo}} = d_{\text{SimCLR}} = d_{\text{SwAV}} = 128$, and $d_{\text{BYOL}} = 256$, as in their original publications [36, 12, 9, 32].

**BYOL details.** Our BYOL implementation uses a momentum annealing starting from $m_{\text{base}} = 0.996$. We minimize the negative cosine similarity in equation (2) of the main paper multiplied by 2 which is equivalent to BYOL’s MSE of $l_2$-normalized vectors [32]. The projection and prediction MLPs have 2 layers (one hidden layer with dimension 4096) and use BN following the original publication [32].

**SimCLR details.** We follow the default implementation [12]. We use a 3-layer projection MLP with a hidden dimension of 2048, ReLU and BN. The loss in equation (1) of the main paper is computed synchronized over the full batch size.
SwAV details. We follow the default implementation [9], using 3 Sinkhorn-Knopp iterations [15] and freezing the prototypes for the first epoch. The Sinkhorn regularization parameter is set to 0.05. As in the default implementation [9], the matrix normalization statistics of the Sinkhorn-Knopp algorithm are computed synchronized over the full training batch. The projection MLP uses ReLU and BN and is identical to the one used in [9], only that we use a 3-layer MLP instead of 2 (ablations are in Table 6 of the main paper).

Encoder details. Our default encoder, \( f_k \), is a R-50 Slow model [20], i.e., a ResNet-50 [39] with a temporal dimension of size \( T \) and sample rate \( \tau \). We perform all ablations with default \( T \times \tau \) of 8×8. We show the architecture in Table 17.

Augmentation details. We perform video decoding and data augmentation using PyTorch’s torchvision package.

We obtain different clips from a video by the following procedure. For the temporal dimension, we randomly sample a clip of \( T \times \tau \) frames from the full-length video, and the input to the ResNet encoder are \( T \) frames subsampled from the raw clip with a stride of \( \tau \); for the spatial dimension, we randomly crop \( 224 \times 224 \) pixels from a video, or its horizontal flip, with a shorter side randomly sampled in [256, 320] pixels [20] (VGG-style [76, 39] spatial cropping, a comparison to Inception-style [80] cropping, which we use for results in §4.5, is given in Table 9 of the main paper).

To each clip, we apply a random horizontal flip, color distortion and Gaussian blur following the SimCLR algorithm are computed synchronized over the full training batch. The projection MLP uses ReLU and BN and is identical to the one used in [9], only that we use a 3-layer MLP instead of 2 (ablations are in Table 6 of the main paper).

| stage | kernels | output sizes \( T \times S^2 \) |
|-------|---------|------------------|
| raw clip | - | \( T \times 224^2 \) |
| data layer | stride \( \tau \), 1 | \( T \times 224^2 \) |
| conv1 | 1×7, 64; stride 1, 2 | \( T \times 112^2 \) |
| pool1 | 1×3 max; stride 1, 2 | \( T \times 56^2 \) |
| res2 | 1×1, 64; 1×3, 64; 1×1, 256 | \( T \times 56^2 \) |
| res3 | 1×1, 128; 1×3, 128; 1×1, 512; 1×1, 256 | \( T \times 28^2 \) |
| res4 | 3×1, 256; 1×3, 256; 1×1, 1024 | \( T \times 14^2 \) |
| res5 | 3×1, 512; 1×3, 512; 1×1, 2048 | \( T \times 7^2 \) |
| pool5 | global average pool | \( 1 \times 1^2 \) |

Table 17. R-50, Slow pathway [20]. The dimensions of kernels are denoted by \( \{T \times S^2, C\} \) for temporal, spatial, and channel sizes. Strides are denoted as \{temporal stride, spatial stride\}. Non-degenerate temporal filters are underlined. Residual blocks are in brackets. Temporal pooling is only performed at the last layer, collapsing spacetime dimensions. By default \( T \times \tau = 8 \times 8 \).

Training augmentation. We use the default training augmentation [20]. We randomly sample a clip of \( T \times \tau \) frames from the full-length video and randomly crop \( 224 \times 224 \) pixels from a video, or its horizontal flip, with a shorter side randomly sampled in [256, 320] pixels.

Inference. Following common practice, in video classification [20], we report 30-view, top-1 classification accuracy on the Kinetics validation set. We uniformly sample 10 clips from a video along its temporal axis. For each clip, we scale the shorter spatial side to 256 pixels and take 3 crops of 256×256 to cover the spatial dimensions. We average the softmax scores for prediction.

B.2. Details: Kinetics Action Classification

Datasets. Kinetics-400 [47] consists of \( \sim 240k \) training videos and 20k validation videos in 400 human action categories. Kinetics-600 [10] has \( \sim 392k \) training videos and 30k validation videos in 600 classes. Kinetics-700 [11] has \( \sim 523k \) training videos and 35k validation videos in 600 classes.

Linear classification protocol. We validate the methods by linear classification on frozen features, following the common protocol in image classification [36]. After unsupervised pre-training on Kinetics, we freeze the features of the encoder and train a linear classifier on top of the last layer features (e.g., pool5 in Table 17). For all ablations in the paper the classifier is trained for 60 epochs (using 100 epochs will increase accuracy by \( \sim 0.2\% \)) using the same cosine schedule as for pre-training (Sec. B.1) with a base learning rate of \( \eta = 4.0 \) (10× higher than in pre-training), linear warm-up in the first 8 epochs, and weight decay of 0.
(cf. Table 17) by extending a 2D proposal at a frame into a 3D RoI by replicating it along the temporal axis, followed by application of frame-wise RoIAlign [37] and temporal global average pooling. We set the spatial stride of res5 to 1 (instead of 2), and use a dilation of 2 for its filters [20]. This increases the spatial resolution of res5 by 2×. The RoI features are then max-pooled and fed to a per-class sigmoid classifier for prediction.

**Training.** For direct comparison, the training procedure and hyper-parameters for AVA follow [20] without modification. The network weights are initialized from the Kinetics models and we use step-wise learning rate decay, that is reduced by 10× after 16, 24 and 28 epochs. We train for 32 epochs on ∼211k data, with linear warm-up [30] for the first 5 epochs and use a weight decay of 10−7, as in [20]. For 8 GPU training, we use a batch-size of 64, a learning rate of 0.05 for the supervised pre-trained Kinetics models and 0.3 for the unsupervised ones, as this gives the best result for each of them.

The region proposal extraction also follows [20] and is summarized here for completeness. Our region proposals are computed by an off-the-shelf person detector, i.e., that is not jointly trained with the action detection models. We adopt a person-detection model trained with Detectron [25]. It is a Faster R-CNN with a ResNeXt-101-FPN backbone. It is pre-trained on ImageNet and the COCO human keypoint images [55]. We fine-tune this detector on AVA for person (actor) detection. The person detector produces 93.9 AP@50 on the AVA validation set. Then, the region proposals for action detection are detected person boxes with a confidence of > 0.8, which has a recall of 91.1% and a precision of 90.7% for the person class.

**Inference.** We perform inference on a single clip with 8 frames sampled with stride 8 centered at the frame that is to be evaluated.

**B.4. Details: Charades Action Classification**

**Dataset.** Charades [75] has ∼9.8k training videos and 1.8k validation videos in 157 classes in a multi-label classification setting of longer activities spanning ∼30 seconds on average. Performance is measured in mean Average Precision (mAP).

**Training.** For Charades, we fine-tune the Kinetics models, but extend their duration by 2× (T × τ = 16 × 8) to account for the long-term nature of the dataset. This increase accuracy of all models by ∼3 mAP. Our training augmentation is the same as in §B.2. A per-class sigmoid output is used for multi-class prediction. We train for 60 epochs using a batch size of 64 and a base learning rate of 0.2 for supervised pre-trained Kinetics models and 0.3 for the unsupervised ones, as this gives the best result for each of them.

**References**

[1] Pulkit Agrawal, João Carreira, and Jitendra Malik. Learning to see by moving. In Proc. ICCV, pages 37–45. IEEE, 2015.

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In Proc. NeurIPS, 2020.
[3] Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In Proc. NeurIPS, 2020. 2, 8

[4] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In Proc. ICCV, 2017. 2

[5] Relja Arandjelović and Andrew Zisserman. Objects that sound. In Proc. ECCV, 2018. 2

[6] Suzanna Becker. Learning temporally persistent hierarchical representations. In Proc. NeurIPS, 1997. 2

[7] 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 Proc. CVPR, 2020. 8

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In Proc. ECCV, 2018. 2, 8

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882, 2020. 1, 2, 3, 4, 7, 11, 12

[10] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloé Hillier, and Andrew Zisserman. A short note about kinetics-600. arXiv preprint arXiv:1808.01340, 2018. 9, 12

[11] João Carreira, Eric Noland, Chloé Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987, 2019. 9, 12

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020. 1, 2, 3, 4, 6, 7, 11, 12

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In Proc. NeurIPS, 2020. 11

[14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020. 6, 11, 12

[15] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Proc. NeurIPS, 2013. 12

[16] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. DynamoNet: Dynamic Action and Motion Network. In Proc. ICCV, 2019. 2

[17] Carl Doersch, Abhinav Gupta, and Alexei Efros. Unsupervised visual representation learning by context prediction. In Proc. ICCV, 2015. 2

[18] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE PAMI, 38(9):1734–1747, Sept 2016. 1, 2

[19] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. PySlowFast. https://github.com/facebookresearch/slowfast, 2020. 5

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In Proc. ICCV, 2019. 2, 4, 5, 7, 8, 12, 13

[21] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In Proc. ICCV, 2017. 2

[22] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In Proc. CVPR, 2018. 2

[23] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017. 8

[24] Deepthi Ghadiyaram, Matt Feisli, Du Tran, Xuetin Yan, Heng Wang, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In Proc. CVPR, 2019. 4, 5, 8, 9

[25] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 13

[26] R. B. Girshick. Fast R-CNN. In Proc. ICCV, 2015. 12

[27] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 315–323, 2011. 11

[28] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. arXiv preprint arXiv:2003.07990, 2020. 2

[29] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In Proc. ICCV, 2015. 2

[30] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training ImageNet in 1 hour. arXiv:1706.02677, 2017. 13

[31] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalinski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The “Something Something” video database of ungrounded actions. In Proc. CVPR, 2017. 1, 4, 7, 13

[32] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheishlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In NeurIPS, 2020. 1, 2, 3, 4, 7, 10, 11

[33] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatiotemporally localized atomic visual actions. In Proc. CVPR, 2018. 1, 4, 7, 12

[34] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In Workshop on Large Scale Holistic Video Understanding, ICCV, 2019. 2
[71] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. arXiv preprint arXiv:2008.03800, 2020. 2, 8

[72] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proc. NeurIPS, 2016. 12

[73] Pierre H Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, et al. Byol works even without batch statistics. arXiv preprint arXiv:2010.10241, 2020. 10

[74] Pierre Sermanet et al. Time-contrastive networks: Self-supervised learning from video. In Proc. Intl. Conf. on Robotics and Automation, 2018. 2

[75] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In ECCV, 2016. 1, 4, 7, 13

[76] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. ICLR, 2015. 7, 12

[77] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012. 2, 4, 13

[78] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In Proc. ICML, 2015. 2

[79] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. arXiv preprint arXiv:1906.05743, 2019. 2

[80] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proc. CVPR, 2015. 7, 12

[81] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Proc. ECCV, 2020. 2

[82] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In Proc. CVPR, 2018. 5

[83] Aáron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018. 2, 3

[84] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabelled video. In Proc. CVPR, 2016. 2

[85] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In Proc. ICCV, 2015. 2

[86] Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning. In Proc. ICCV, 2017. 2

[87] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In Proc. CVPR, 2019. 2

[88] Laurenz Wiskott and Terrence Sejnowski. Slow feature analysis: Unsupervised learning of invariances. In Neural Computation, 2002. 2, 5

[89] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. In Proc. CVPR, volume abs/1805.01978, 2018. 1, 2

[90] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. In Proc. ECCV, 2018. 5

[91] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In Proc. CVPR, 2019. 2

[92] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. arXiv preprint arXiv:2006.15489, 2020. 2, 8

[93] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888, 2017. 10, 11

[94] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Proc. ECCV, 2016. 2

[95] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In Proc. ICCV, 2019. 2