Learning from Untrimmed Videos:
Self-Supervised Video Representation Learning with Hierarchical Consistency

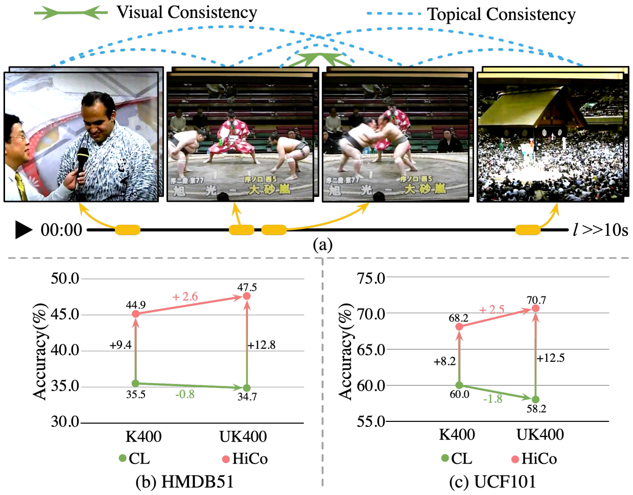
Zhiwu Qing[1] Shiwei Zhang[2*] Ziyuan Huang[3] Yi Xu[4] Xiang Wang[1]
Mingqian Tang[2] Changxin Gao[1*] Rong Jin[2] Nong Sang[1]
[1]Key Laboratory of Image Processing and Intelligent Control
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Alibaba Group [3]ARC, National University of Singapore [4]Dalian Unversity of Technology
{qzw, wxiang, cgao, nsang}@hust.edu.cn
{zhangjin.zsw, mingqian.tmq, jinrong.jr}@alibaba-inc.com
ziyuan.huang@u.nus.edu yxu@dlut.edu.cn

Abstract

Natural videos provide rich visual contents for self-supervised learning. Yet most existing approaches for learning spatio-temporal representations rely on manually trimmed videos, leading to limited diversity in visual patterns and limited performance gain. In this work, we aim to learn representations by leveraging more abundant information in untrimmed videos. To this end, we propose to learn a hierarchy of consistencies in videos, i.e., visual consistency and topical consistency, corresponding respectively to clip pairs that tend to be visually similar when separated by a short time span and share similar topics when separated by a long time span. Specifically, a hierarchical consistency learning framework ***HiCo*** *is presented, where the visually consistent pairs are encouraged to have the same representation through contrastive learning, while the topically consistent pairs are coupled through a topical classifier that distinguishes whether they are topic-related. Further, we impose a gradual sampling algorithm for proposed hierarchical consistency learning, and demonstrate its theoretical superiority. Empirically, we show that not only HiCo can generate stronger representations on untrimmed videos, it also improves the representation quality when applied to trimmed videos. This is in contrast to standard contrastive learning that fails to learn appropriate representations from untrimmed videos.*

1. Introduction

Self-supervised learning is of crucial importance in computer vision, and has shown remarkable potential in learning powerful spatio-temporal representations using unlabelled videos. Current state-of-the-art approaches on unsupervised video representation learning are typically based on the contrastive learning framework [13, 44, 45], which encourages the representations of the clips from the same video to be close and those from different videos to be as far away from each other as possible [7, 21]. In most approaches, they are trained on manually trimmed videos such as Kinetics-400 [5]. However, it is labor-intensive and time-consuming to collect such a large-scale trimmed video dataset, and the



(b) HMDB51

(c) UCF101

Figure 1. (a) An example of untrimmed video with the interview, competition and the stadium of *Sumo Wrestling*. It shows the hierarchical consistency present in untrimmed videos. As can be seen, clips with short temporal distance share similar visual elements, while clips with long temporal distance, despite their dissimilar visual contents, share a same topic. (b, c) Linear evaluation of conventional contrastive learning (CL), *i.e*., SimCLR [7], and HiCo on HMDB51 [27] and UCF101 [49], with pretraining respectively on the original (trimmed) and untrimmed version of Kinetics-400 [5].

*Corresponding authors.
Project page: https://hico-cvpr2022.github.io/.

trimming process may also bring in certain human bias into the data. In contrast, natural videos carry more abundant and diverse visual contents, and they are easier to obtain. Hence, this work sets out to exploit the natural *untrimmed videos* for video representation learning.

Directly learning generalized and powerful representations from untrimmed videos is not a trivial problem, as empirical results both in Fig. 1(b, c) and in [13](Tab.4 and Tab.6) demonstrate that directly applying contrastive learning on untrimmed videos yields *worse* representations than on trimmed videos. One possible reason is that the temporally-persistent hypothesis [13] followed by the standard video contrastive learning framework and verified on trimmed videos is no longer sufficient for untrimmed videos. Ideally, the temporally-persistent hypothesis learns an invariant representation for all the clips in the video. This may be plausible for trimmed videos, and even for clips with short temporal distance in untrimmed ones, where a certain level of visual similarity or *visual consistency* exists. Yet it could be *overly strict* for temporally distant clips in untrimmed videos with less or no visual consistency exists, since they are only related by the same topic, *i.e.,* they are *topically consistent*. In fact, we spot a hierarchical relation between the two consistencies existing in untrimmed videos. Specifically, visually consistent pairs are always topically consistent, while topically consistent pairs are not necessarily visually consistent. An example of the hierarchical consistency is visualized in Fig. 1(a).

In this paper, we present a novel framework for learning strong representations from *untrimmed videos*. By exploiting the hierarchical consistencies existing in untrimmed videos, *i.e.,* the visual consistency and the topical consistency, our framework **HiCo** for Hierarchical Consistency learning can leverage the more abundant semantic patterns in natural videos. We design two hierarchical tasks, respectively for learning the two consistencies. For visually consistent learning, we apply standard contrastive learning on clips with a small maximum temporal distance, and encourage temporally-invariant representations. For topical consistency learning, we propose a topic prediction task, instead of a strict invariant mapping, the representations are only required to group different topics. Considering the hierarchical nature of consistencies, we also include the visually consistent pairs in topical consistency learning, while exclude topically consistent pairs for visual consistency learning. Due to the complexity of the hierarchical tasks, we further introduce a gradual sampling that gradually increases the training difficulty for positive pairs to help optimization and improve generalization, which we show its superior both theoretically and empirically.

Extensive experiments on multiple downstream tasks show that employing HiCo can learn a strong and generalized video representation from untrimmed videos, with a convincing gap of 12.8% and 12.5% on the downstream action recognition task respectively on HMDB51 [27] and UCF101 [49] compared with the standard contrastive learning. We also demonstrate the capability of HiCo to learn a better representation from trimmed videos.

2. Related Works

Long video understanding. The existing attempts for long video understanding is mainly based on supervised learning. *Shot or event boundary detection* approaches [2,17,47, 48,50,53] aim to detect shot transitions or event boundaries in untrimmed videos. Among them, the former is caused by manual editing, and the latter is semantically-coherent. For *temporal action localization*, existing works [15,32,33, 46, 66] attempt to distinguish action instances from unrelated complex backgrounds by modeling temporal relationships in untrimmed videos. Although the complex temporal structures in videos bring challenges for these tasks, there are many *video classification* methods [31, 36, 41, 52, 54, 62, 67, 70] aggregate long-range temporal context to augment the predictions and achieve remarkable performance. Unfortunately, these excellent supervised methods can not be transferred to self-supervised learning. In this work, we try to leverage the inherent temporal structure in untrimmed videos for self-supervised video representation learning.

Self-supervised image representation learning. To avoid the labor-intensive annotation process, a wide range of self-supervised approaches have been proposed to exploit unlabeled data. Early methods mainly design different pretext tasks, including color restoration [72], image context reconstruction [42] and solving jigsaw puzzles [11, 38], *etc*. Recently, contrastive learning based on instance discrimination has shown great potential in this field [7,8,21,22,39, 56, 63]. The main idea of contrastive learning is to train a transformation-invariant network.

Self-supervised video representation learning. Existing self-supervised video representation learning approaches can be divide into three groups: designing different pretext tasks, applying contrastive learning and combining the both. Pretext task based methods exploit the inherent structures naturally existing in videos to supervise the networks, such as speed perception [3, 61, 69], order prediction [14, 30, 37, 65, 74], temporal transformation discrimination [26], motion estimation [24, 60], and future prediction [9, 18, 35, 51, 59]. Contrastive learning related works [10, 13, 40, 44, 45] are mainly extended from image paradigm, and explore various spatio-temporal transformations for videos. It is worth noting that existing state-of-the-art methods are almost all based on the contrastive learning framework. Further, there exist approaches to combine contrastive learning and temporal pretext tasks into a multi-task learning framework [1, 25, 29, 55], which enable the temporal exploration ability for contrastive learn-

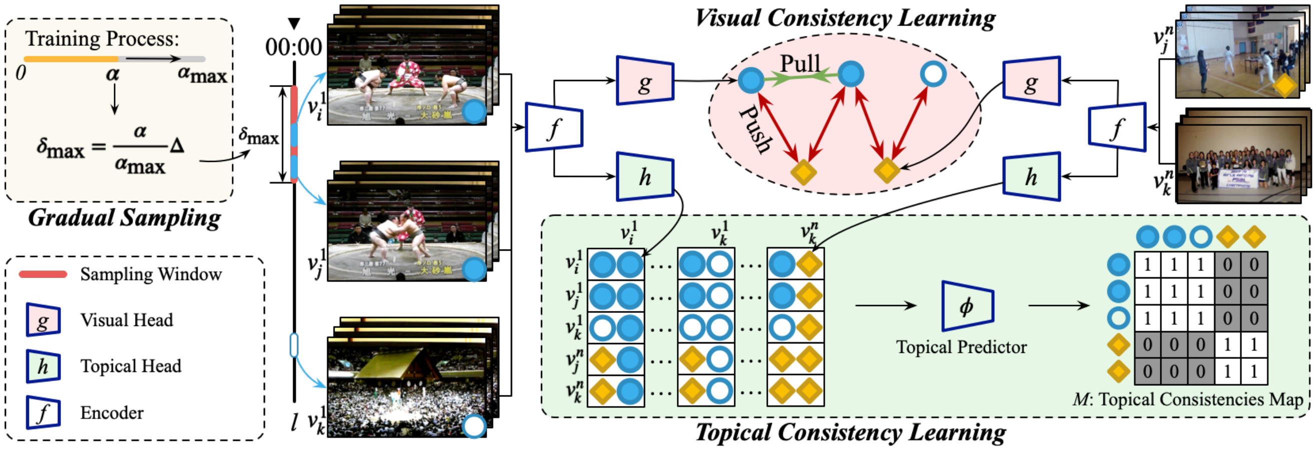


Figure 2. The overall framework of HiCo. HiCo contains three parts, including Visual Consistency Learning (VCL), Topical Consistency Learning (TCL), and Gradual Sampling (GS). VCL is based on standard contrastive learning to map a shared visual embedding for visually consistent pairs. TCL learns a topical predictor to discriminate the topical consistency between any two clips. The purpose of GS is to enhance both VCL and TCL by controlling the difficulty of training clips in each video.

ing and can further improve the video representations. Although previous methods have made significant progress for self-supervised video representation learning, they mostly rely on the curated videos that are manually trimmed beforehand and ignore the rich visual patterns embedded in original untrimmed videos. In contrast, HiCo is a first attempt that focuses on self-supervised learning in untrimmed videos and enjoys both short-range and long-range temporal contexts simultaneously, as far as we know.

3. Hierarchical Consistency Learning

The main difference between the untrimmed videos and trimmed ones lies in the video length. For trimmed videos, any two random clips are likely to be visually similar since the temporal distance between two clips is always small. However, for untrimmed videos, randomly sampled clip pairs could have a long temporal distance, making them only topically related, with low visual similarity. On the other hand, clip pairs with short temporal distance could still be viewed as two clips sampled from trimmed videos, which share a high visual similarity. Hence, we divide the relations between the clip pairs into a hierarchy: **(i)** for clip pairs with short temporal distance with high visual similarity, we define their relationship as *visually consistent;* **(ii)** for clip pairs with long temporal distance that may be only topic-related but visually dissimilar, we define their relationship as *topically consistent.* Corresponding to this, we propose two hierarchical tasks to learn from the hierarchical consistencies, respectively visual consistency learning (VCL in Sec. 3.1) and topical consistency learning (TCL in Sec. 3.2). Considering the complexity of the hierarchical tasks, we further propose a novel Gradual Sampling strategy to improve both VCL and TCL, and also provide theoretical analysis for its effectiveness. Combined together, we present our overall framework HiCo in Fig. 2.

3.1. Visual Consistency Learning

We learn visual consistency using the contrastive learning method SimCLR [7]. When applying contrastive learning for videos, it learns to map different clips from the same video (*i.e.*, positive pairs) closer and repel clips from different videos (*i.e.*, negative pairs). Specifically, in a minibatch of N videos, it samples two clips v_i and v_j from each video and thus generates $2N$ views with independent data augmentations. After one latent vector $\mathbf{z}$ is extracted for each view through a backbone and a projection layer, the loss for the contrastive learning is formulated as:

$$\mathcal{L}_{\text{CL}} = \frac{1}{2N}\sum_{n=1}^{N}[\ell(2n-1,2n)+\ell(2n,2n-1)] \ , \qquad (1)$$

where $\ell(i,j)$ denotes the loss between two paired samples. Given the cosine similarity $s_{i,j}$ between the representations $\mathbf{z}_i$ and $\mathbf{z}_j$, where $\{\mathbf{z}_i, \mathbf{z}_j\} = g(f(v_i, v_j))$ with f being the video backbone and g being the contrastive projection head, $\ell(i,j)$ can be calculated as:

$$\ell(i,j) = -\log\frac{\exp(s_{i,j}/\tau)}{\sum_{n=1}^{2N}\mathbb{1}_{[n\neq i]}\exp(s_{i,n}/\tau)} \ , \qquad (2)$$

where τ represents the temperature and $\mathbb{1}_{[n\neq i]}$ equals 1 if $n \neq i$, otherwise 0.

Since random sampling may yield v_i and v_j with low visual similarities in untrimmed videos, we further limit the maximum temporal distance for the clip pairs to learn the visual consistency. Formally, the temporal distance $\delta(v_i, v_j)$ between v_i and v_j is calculated and limited as:

$$\delta(v_i, v_j) = |c_i - c_j| < \delta_{\max} \ , \qquad (3)$$

where c_i and c_j is the time step of the central frame in v_i and v_j, and δ_{max} denotes the maximum distance between two sampled clips for visual consistency learning. To guarantee the visual consistency between v_i and v_j, δ_{max} should be significantly smaller than the video duration l, *i.e.,* $0 \leq \delta_{\text{max}} \ll l$.

3.2. Topical Consistency Learning

In general, the distant clips in untrimmed videos may be visually dissimilar but share the same topics, which is shown in the example in Fig. 1 (a). Although the scenes of interviews and stadium share little visual similarity with the competition, they all belong to the same topic of *Sumo Wrestling*. Hence, to fully exploit the visual diversities in untrimmed videos, we propose to learn from this topical consistency, which is overlooked in previous approaches.

Formally, to learn topical consistency, we additionally randomly sample another clip v_k from the entire video, which is not necessarily visually consistent to v_i and v_j, but topically consistent to them. However, due to the potential significant visual variations, it would be unreasonable for the topically consistent pairs to learn an invariant mapping. Therefore, we relax this strict constraint by (a) only introducing the v_k as the negative sample for other videos in VCL; and (b) designing a learnable predictor to distinguish whether the input pairs are topically consistent, *i.e.,* whether they belong to the same video.

With v_k in the negative sample pool, the loss between visually consistent pairs $\ell(i,j)$ are now calculated as follows:

$$\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{n=1}^{3N} \mathbb{1}_{[n \neq i,k]} \exp(s_{i,n}/\tau)} . \tag{4}$$

For topic prediction, we first obtain the topical representations $\{\mathbf{t}_i, \mathbf{t}_j, \mathbf{t}_k\}$ for the sampled clips $\{v_i, v_j, v_k\}$ by the encoder $f(\cdot)$ and a topical project head $h(\cdot)$, *i.e.,* $\{\mathbf{t}_i, \mathbf{t}_j, \mathbf{t}_k\} = h(f(\{v_i, v_j, v_k\}))$. Given the N videos with $3N$ clips in each mini-batch, the topical representations for all videos are combined to form a pair-wise feature set $\mathbf{U}$:

$$\mathbf{U} = \left\{ \begin{matrix} \mathbf{t}_i^1 \oplus \mathbf{t}_i^1, & \mathbf{t}_i^1 \oplus t_j^1, & \cdots & \mathbf{t}_i^1 \oplus \mathbf{t}_j^N, & \mathbf{t}_i^1 \oplus \mathbf{t}_k^N, \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{t}_k^N \oplus \mathbf{t}_i^1, & \mathbf{t}_k^N \oplus \mathbf{t}_j^1, & \cdots & \mathbf{t}_k^N \oplus \mathbf{t}_j^N, & \mathbf{t}_k^N \oplus \mathbf{t}_k^N, \end{matrix} \right\}, \tag{5}$$

where the superscript $1...N$ denotes the video index, $\oplus$ denotes the concatenation and $\mathbf{U} \in \mathbb{R}^{3N \times 3N \times 2C_{\text{T}}}$ with C_{T} being the dimension of the topical representation. Finally, the topical consistencies M for these pair-wise clips are estimated by a topical predictor:

$$M = \phi(\mathbf{U}) \in R^{3N*3N}. \tag{6}$$

where the topical predictor $\phi(\cdot)$ is implemented by a Multi-Layer Perceptron (MLP). The supervised label for the topical consistencies M is defined as $G \in R^{3N*3N}$, which indicates whether the pair-wise features share the same topic (*i.e.*, whether they are from the same video). During training, we apply focal loss [34] $\mathcal{F}$ since the number of topically consistent pairs and inconsistent ones are heavily unbalanced. The topic prediction loss is calculated as follows:

$$\mathcal{L}_{\text{TP}} = \frac{1}{\gamma_1} \sum_{G_{i,j}=1} \mathcal{F}(M_{i,j}) + \frac{1}{\gamma_2} \sum_{G_{i,j}=0} \mathcal{F}(1 - M_{i,j}), \tag{7}$$

where γ_1 and γ_2 are the number of positive samples and negative samples. Compared to visual consistency learning, where representations of the same video are encouraged to be identical, topical consistency learning poses a less strict constraint on the representations. Finally, the overall training objective of our HiCo is the sum of the contrastive loss and the topic prediction loss, formulated as $\mathcal{L} = \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{TP}}$.

3.3. Gradual Sampling

Curriculum learning [4] shows that models can learn much better when the training examples are not randomly provided but organized in a meaningful order, from easy examples to the hard ones. It has achieved great success in a wide range of tasks. Recalling that the untrimmed videos usually contain complex temporal contexts, randomly sampling clips unavoidablely generates dissimilar pairs in the early training stage, which can be considered as *hard examples* . Therefore, we bring the spirit of curriculum learning into our HiCo, and propose a simple yet effective strategy to control the difficulty of positive pairs during the training stage, termed as Gradual Sampling.

Specifically, the δ_{max} is no longer a constant, but a function is driven by the current training epoch α:

$$\delta_{\text{max}}(\alpha) = \frac{\alpha}{\alpha_{\text{max}}} \Delta, \tag{8}$$

where α and α_{max} refer to the current training epoch and the total training epoch, respectively. Δ is the upper bound of $\delta_{\text{max}}(\alpha)$, since $\alpha/\alpha_{\text{max}}$ satisfies the condition: $\alpha/\alpha_{\text{max}} \in [0, 1]$, and here both α_{max} and Δ are constants. This gradual sampling can be utilized to sample both visually consistent clips and topically consistent clips.

The $\delta_{\text{max}}(\alpha)$ linearly grows from 0 to Δ, which means we train the network from identical clips (with different data augmentations) and gradually increase the difficulty of positive pairs. This can help improve the video representation generalization, and we will theoretically and experimentally show its superiority. In fact, the gradual sampling in self-supervised learning can be also applied to both trimmed and untrimmed videos.

Theoretical analysis. We provide a theoretical understanding of the proposed gradual sampling strategy by leveraging

For Visual Consistency Learning, even we limit the maximum temporal distance between clips v_i and v_j, the training pairs are still considered *harder* with a large temporal distance.

the generalization analysis, which is common in the literature of learning theory [58]. For the sake of analysis simplicity, we abstract the key points from the strategy, which are more math-friendly. To this end, we divide the training data into two groups, one with small variance (denoted by $\widehat{\mathcal{D}}_s$) and another one with large variance (denoted by $\widehat{\mathcal{D}}_l$). We let their population distributions as $\mathcal{D}_s$ and $\mathcal{D}_l$, respectively. Please note that this partition is for the proof use only, and it is not required in practice. At the early training epochs, the sampled clips are considered as examples with small variance since the sampling window size is small according to Eq.8. While during the later training epochs, the sampled clips could be examples either with large or with small variance due to the larger sampling window size. Let $\mathcal{L}(w)$ be the loss function of the deep learning task that aims to be optimized, where w is the model parameter. Given the output $\widehat{w}$ of an algorithm, the excess risk (ER) is a standard measure of generalization in learning theory [58], whose formulation is $\mathcal{L}(\widehat{w})-\mathcal{L}(w_*)$, where $w_* = \arg\min_w \mathcal{L}(w)$. The main goal is to obtain a solution $\widehat{w}$ as close as to the global optimal w_*. The following informal theorem (please refer to Appendix for its formal version) presents two excess risk bounds (ERB) to theoretically show why Gradual Sampling (GS) based sampling has better generalization than Random Sampling (RS) under some mild assumptions. Due to the space limitation, we include all other details, formal theorem, and proof in *Appendix*.

Theorem 1 (Informal)**.** *Under some mild assumptions, the GS strategy can yield better generalization than the RS strategy. Specifically, we have the following ERB in expectation: (1) for output of RS $\widehat{w}_{rs}$,*

$$\mathcal{L}(\widehat{w}_{rs}) - \mathcal{L}(w_*) \leq O\left(\mathcal{L}(w_0) - \mathcal{L}(w_*)\right);$$

and (2) for output of GS $\widehat{w}_{gs}$,

$$\mathcal{L}(\widehat{w}_{gs}) - \mathcal{L}(w_*) \leq O\left(\log(n)/n + p^2\hat{\Delta}^2\right),$$

where w_0 is the initial solution, here $\hat{\Delta}$ is a measurement of difference between $\mathcal{D}_s$ and $\mathcal{D}_l$, n is the sample size of $\widehat{\mathcal{D}}_s$ and $p \in [0,1]$ is the proportion of $\widehat{\mathcal{D}}_l$ among all training examples.

The result (1) shows that RS did not receive a significant reduction in the objective due to the large variance arising from D_l. On the other hand, the result (2) tells that GS could reduce the objective significantly when n is large and p is small, showing it has better generalization than RS. Please note that by appropriately selecting $\widehat{\mathcal{D}}_s$ and $\widehat{\mathcal{D}}_l$, n can be large and p is small enough, while theoretically the constant $\mathcal{L}(w_0) - \mathcal{L}(w_*)$ could be very large in general.

4. Experiments

Pre-training dataset. *Kinetics-400* [5] *(K400)* contains 240k trimmed videos, and each video lasts about 10 seconds. Since these short videos are trimmed from long videos, we recollect their original versions as our untrimmed video dataset, which we call *untrimmed Kinetics-400 (UK400)*. Because many original videos are unavailable now, our UK400 dataset only contains 157k untrimmed videos for pre-training. *HACS* [73] is a large-scale dataset for temporal action localization, which contains 37.6k long untrimmed videos for training.

Pre-training settings. We choose SimCLR [7, 44] as basic contrastive learning framework, and adopt three frequently-used networks as encoder $f(\cdot)$, including S3D-G [64], R(2+1)D-10 [57] and R3D-18 [20]. More training details about pre-training please refer to *Appendix*.

Evaluations. We evaluate the representations learned by HiCo on three different downstream tasks, including action recognition, video retrieval and temporal action localization. Among them, action recognition and video retrieval are performed on two datasets: UCF101 [49] and HMDB51 [27]. For temporal action localization, we employ ActivityNet [12] as evaluation dataset. Please refer to *Appendix* for more fine-tuning settings.

Note. In this section, unless otherwise specified, 'FT/LFT' refers to fully fine-tuning/linear fine-tuning. 'VCL', 'TCL' and 'GS' are Visual Consistency Learning, Topical Consistency Learning and Gradual Sampling, respectively. Symbols '✓' and '✗' respectively indicate 'Yes' and 'No'.

4.1. Ablation Study

Importance of proposed VCL, TCL, and GS. Tab. 1 ablates different components of HiCo. From the results, we can find that: first, VCL can improve standard contrastive learning on both datasets, which demonstrates visually consistent short-range clips can enhance the representation quality; second, TCL alone is relatively weaker than VCL, but they are complementary to each other. Combining VCL and TCL can respectively gain 7.2% and 10.6% on HMDB51 and UCF101 when pre-trained on UK400, as shown in Tab. 1 (a); third, GS can significantly improve VCL, TCL and their combination, especially it improves 5.6% (41.9% *vs.* 47.5%) on HMDB51 with UK400 pre-training. Meanwhile, a similar trend is observed with HACS pre-training. These results demonstrate the effectiveness of each component of HiCo.

Parameter sensitivity analysis for the upper bound of $\delta_{\max}(\alpha)$, *i.e.*, Δ. Tab. 2(a) presents results for varying Δ. It shows that the best performance is obtained when Δ=1.0s. Note that, when Δ is set to 0s, the performances respectively drop 6.1% and 5.1% on HMDB51 and UCF101, since all training examples are identical pairs without temporal variance, which declines the generalization. Conversely, a

When the window size is small, the sampled clips are usually similar.

PT.	VCL	TCL	GS	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
HACS	✗	✗	✗	42.9/29.5	75.6/54.9
	✓	✗	✗	42.7/33.8	76.6/57.9
	✗	✓	✗	42.6/31.5	74.9/55.9
	✓	✓	✗	43.9/**35.6**	75.2/**64.8**
UK400	✗	✗	✗	45.1/34.7	74.7/58.2
	✓	✗	✗	47.9/37.8	77.4/65.2
	✗	✓	✗	46.1/34.8	77.2/60.9
	✓	✓	✗	50.5/**41.9**	77.7/**68.8**

(a)

PT.	VCL	TCL	GS	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
HACS	✗	✗	✓	43.8/31.0	75.3/57.4
	✓	✗	✓	48.7/37.7	76.2/63.0
	✗	✓	✓	45.2/32.0	75.3/58.7
	✓	✓	✓	**51.8/41.6**	**77.6/67.6**
UK400	✗	✗	✓	46.1/33.9	76.8/59.3
	✓	✗	✓	51.2/41.8	78.5/67.2
	✗	✓	✓	49.9/36.1	76.7/62.4
	✓	✓	✓	**54.1/47.5**	**79.6/70.7**

(b)

U	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
Uni.	52.4/45.7	79.3/69.0
Bi.	**54.1/47.5**	**79.6/70.7**

(c)

v_k	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
✗	48.8/40.5	77.0/68.0
✓	50.5/**41.9**	77.7/**68.8**

(d)

Table 1. Ablation studies for HiCo with S3D-G. (a, b) Evaluating VCL and TCL both with and without GS. 'PT.' refers to 'Pre-training dataset'. (c) Bidirectional (Bi.) and unidirectional (Uni.) concatenation in $\mathbf{U}$. (d) Whether v_k is in the negative sample pool of VCL.

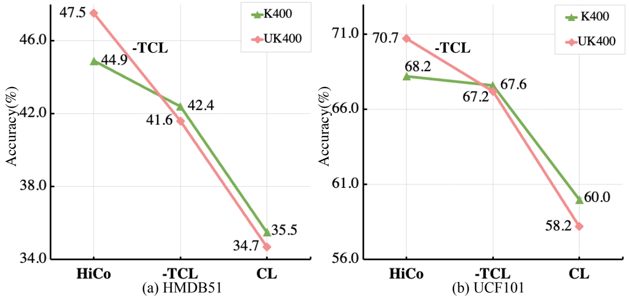
Δ(s)	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
0.0	51.9/41.4	80.4/65.6
0.5	54.9/46.1	79.5/69.4
1.0	54.1/**47.5**	79.6/**70.7**
2.0	51.7/44.9	77.5/69.5
4.0	52.2/43.5	79.6/68.8

(a)

Dis.(s)	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
0	51.5/43.6	78.7/66.8
10	53.3/45.6	80.1/69.4
50	55.1/45.1	78.5/69.7
100	53.1/45.8	80.2/70.3
$+\infty$	54.1/**47.5**	79.6/**70.7**

(b)

Table 2. Parameter sensitivity analysis. All experiments are conducted on UK400 with S3D-G. (a) The upper bound of $\delta_{\max}(\alpha)$, *i.e.*, Δ. (b) The temporal distance of topical pairs.



(a) HMDB51 (b) UCF101

Figure 3. Removing TCL from HiCo. We pre-train S3D-G on K400 and UK400, and visualize the linear evaluations.

large Δ may introduce dissimilar pairs with large visual variance, which may increase optimization difficulty and hence hurts the learned representations.

Impact of the distance between topical pairs. Increasing the distance between topical pairs can introduce more temporal diversities, and the results under different settings are shown in Tab. 2(b). We observe that the performance increases with a larger distance, *e.g.*, increasing the distance from 0 to $+\infty$ can lead to around 3.8% gains on both datasets. The results show that HiCo can effectively leverage the rich visual patterns from long-range topical pairs.

General applicabilities. Tab. 3 explores the generalities of HiCo with different backbones and datasets. We can observe that HiCo can significantly boost performance, concerning both FT and LFT, from the baseline under all settings. Note that the performance of baseline pre-trained on UK400 is lower than that on K400, while HiCo can gain around 2.5% with UK400 pre-training. To further understand the reason behind this, TCL is removed from HiCo in Fig. 3. We can observe that the representations pre-trained on K400 are still stronger than UK400, similar to standard contrastive learning. However, when integrating TCL, the performance pre-trained on UK400 surpasses K400 by 2.6% on HMDB51, fully demonstrating that TCL can help to utilize the diverse temporal contexts in untrimmed videos

Pretrain	Backbone	HiCo	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
HACS	S3D-G	✗	42.9/29.5	75.6/54.9
		✓	**51.8/41.6**	**77.6/67.6**
	R(2+1)D-10	✗	47.7/35.7	81.3/61.7
		✓	**53.1/43.7**	**81.9/71.3**
	R3D-18	✗	43.5/32.8	72.8/57.8
		✓	**49.5/43.3**	**76.1/65.2**
UK400	S3D-G	✗	45.1/34.7	74.7/58.2
		✓	**54.1/47.5**	**79.6/70.7**
	R(2+1)D-10	✗	47.4/32.0	80.7/57.4
		✓	**50.9/39.9**	**82.1/67.7**
	R3D-18	✗	44.4 /40.0	76.5/65.5
		✓	**47.7/46.3**	**77.8/70.7**
K400	S3D-G	✗	46.2/35.5	76.0/60.0
		✓	**53.0/44.9**	**79.0/68.2**

Table 3. HiCo with different datasets and backbones.

	V.C.Pairs	T.C.Pairs	HMDB51 (FT/LFT)	UCF101 (FT/LFT)
(a)	$\mathcal{L}_{\text{CL}}$	None	47.9/37.8	77.4/65.2
(b)	$\mathcal{L}_{\text{TP}}$	None	45.5/19.7	76.9/27.3
(c)	$\mathcal{L}_{\text{CL}}$	$\mathcal{L}_{\text{CL}}$	46.9/36.0	76.1/62.6
(d)	$\mathcal{L}_{\text{TP}}$	$\mathcal{L}_{\text{TP}}$	49.3/24.7	77.5/38.9
(e)	$\mathcal{L}_{\text{CL}}$+$\mathcal{L}_{\text{TP}}$	$\mathcal{L}_{\text{TP}}$	50.5/41.9	77.7/68.8
(f)	$\mathcal{L}_{\text{CL}}$+GS	None	51.2/41.8	78.5/67.2
(g)	$\mathcal{L}_{\text{TP}}$+GS	None	47.6/21.3	77.0/29.1
(h)	$\mathcal{L}_{\text{CL}}$+$\mathcal{L}_{\text{TP}}$+GS	None	52.3/43.5	77.9/65.7
(i)	$\mathcal{L}_{\text{CL}}$+$\mathcal{L}_{\text{TP}}$	$\mathcal{L}_{\text{TP}}$+GS	50.1/41.9	78.9/69.7
(j)	$\mathcal{L}_{\text{CL}}$+$\mathcal{L}_{\text{TP}}$+GS	$\mathcal{L}_{\text{TP}}$+GS	**54.1/47.5**	**79.6/70.7**

Table 4. Ablation studies on loss functions. 'V.C.Pairs' and 'T.C.Pairs' are visually consistent pairs and topically consistent pairs, respectively.

Method	Backbone	Depth	Pretrain	PT Res.	FT Res.	Freeze	UCF101	HMDB51
CVRL [44]	R3D	50	Kinetics-400	16×224^2	32×224^2	✓	89.8	58.3
CCL [28]	R3D	18	Kinetics-400	8×112^2	8×112^2	✓	52.1	27.8
MLRep [43]	R3D	18	Kinetics-400	16×112^2	16×112^2	✓	63.2	33.4
FAME [10]	R(2+1)D	10	Kinetics-400	16×112^2	16×112^2	✓	72.2	42.2
CoCLR [19]	S3D	23	Kinetics-400	32×128^2	32×128^2	✓	74.5	46.1
HiCo(Ours)	R3D	18	HACS	8×112^2	16×112^2	✓	72.8	45.2
HiCo(Ours)	S3D-G	23	Kinetics-400	16×112^2	16×112^2	✓	75.7	52.3
HiCo(Ours)	R3D	18	UKinetics-400	8×112^2	16×112^2	✓	77.6	52.1
HiCo(Ours)	R(2+1)D	10	Kinetics-400	16×112^2	16×112^2	✓	76.7	49.1
HiCo(Ours)	R(2+1)D	10	UKinetics-400	16×112^2	16×112^2	✓	**78.1**	50.1
HiCo(Ours)	S3D-G	23	UKinetics-400	16×112^2	16×112^2	✓	77.9	**57.6**
VCLR [29]	R2D	50	Kinetics-400	3×224^2	$N/A \times 224^2$	✗	85.6	54.1
ρSimCLR [13]	R3D	50	Kinetics-400	8×224^2	8×224^2	✗	88.9	-
CVRL [44]	R3D	50	Kinetics-400	16×224^2	32×224^2	✗	92.2	66.7
ρBYOL [13]	R3D	50	Kinetics-400	16×224^2	16×224^2	✗	95.5	73.6
VCLR [29]	R3D	18	HACS	$N/A \times 224^2$	$N/A \times 224^2$	✗	67.2	49.3
RSPNet [6]	R3D	18	Kinetics-400	16×112^2	16×112^2	✗	81.1	44.6
MLRep [43]	R3D	18	Kinetics-400	16×112^2	16×112^2	✗	79.1	47.6
ASCNet [23]	R3D	18	Kinetics-400	16×112^2	16×112^2	✗	80.5	52.3
VideoMoCo [40]	R(2+1)D	10	Kinetics-400	32×112^2	32×112^2	✗	78.7	49.2
SRTC [71]	R(2+1)D	10	Kinetics-400	16×112^2	16×112^2	✗	82.0	51.2
FAME [10]	R(2+1)D	10	Kinetics-400	16×112^2	16×112^2	✗	84.8	53.5
SpeedNet [3]	S3D-G	23	Kinetics-400	64×224^2	64×224^2	✗	81.1	48.8
RSPNet [6]	S3D-G	23	Kinetics-400	64×224^2	64×224^2	✗	89.9	59.6
HiCo(Ours)	R3D	18	HACS	8×112^2	16×112^2	✗	77.0	56.2
HiCo(Ours)	S3D-G	23	Kinetics-400	16×112^2	16×112^2	✗	83.2	56.3
HiCo(Ours)	R3D	18	UKinetics-400	8×112^2	16×112^2	✗	87.2	63.7
HiCo(Ours)	R(2+1)D	10	Kinetics-400	16×112^2	16×112^2	✗	85.3	57.9
HiCo(Ours)	R(2+1)D	10	UKinetics-400	16×112^2	16×112^2	✗	86.5	55.6
HiCo(Ours)	S3D-G	23	UKinetics-400	16×112^2	16×112^2	✗	83.6	60.4
HiCo(Ours)	S3D-G	23	UKinetics-400	16×112^2	32×224^2	✗	**91.0**	**66.5**

Table 5. Comparison to other state-of-the-art methods on the action recognition task. Where 'Freeze' indicates freezing the parameters in backbones. 'UKinetics-400' is untrimmed Kinetics-400 dataset. 'PT Res.' and 'FT Res.' are sptial-temporal resolutions in pre-training and fine-tuning, respectively. 'Grey fonts' refers to the backbones different from HiCo.

to learn powerful representations.

Analysis of the loss function. Tab. 4 analyzes the hierarchical properties in HiCo from loss perspective. We have several observations. *(i)* Using $\mathcal{L}_{\text{TP}}$ alone is weaker than $\mathcal{L}_{\text{CL}}$. However, when combining $\mathcal{L}_{\text{TP}}$ and $\mathcal{L}_{\text{CL}}$, $\mathcal{L}_{\text{TP}}$ can further absorb the useful information from the topical pairs and improve the accuracy, *e.g.*, 36.0% *vs.* 41.9% on HMDB51, refer to (c, e). This shows the superiority of our proposed hierarchical learning architecture. *(ii)* From (b, d), we observe significant improvement by introducing topical pairs, which again confirms the importance of temporal diversities. *(iii)* From (a, f) and (b, g), separated $\mathcal{L}_{\text{CL}}$ and $\mathcal{L}_{\text{TP}}$ can be promoted by GS. Further comparing (e) and (j), even with extra topical pairs, GS can also strength the combined $\mathcal{L}_{\text{CL}}$ and $\mathcal{L}_{\text{TP}}$. Especially on HMDB51, integrating GS can improve accuracy by 5.6%. These experiments demonstrate the complementarity between visual and topical pairs and the effectiveness of GS from the loss perspective.

More explorations. *(i)* Tab. 1(c) reports two different concatenating ways for pair-wise topical features in the feature set $\mathbf{U}$. The unidirectional one shows weaker performance, since the bidirectional setting can provide more expert prior; that is, topical consistencies are unrelated to feature orders. *(ii)* Tab. 1(d) explores the necessity of incorporating v_k into the negative pool for visually consistent pairs (*i.e.* v_i and v_j). Although v_k may be visually dissimilar with v_i and v_j, it can also provide extra supervision signals for VCL and improve generalization.

4.2. Evaluation on action recognition task

Tab. 5 compares HiCo with other state-of-the-art methods. We list the relevant settings in details for fair comparisons, including network architectures and training resolutions. From the table, we draw the following conclusions. First, in terms of linear evaluation, HiCo notably outperforms existing methods under similar settings. HiCo surpasses CoCLR [19] by 3.4% and 11.5% on UCF101 and HMDB51, respectively, even they use more frames and optical flow in pre-training. Due to the use of a deeper network and larger resolution in CVRL [44], HiCo achieves

Method	Backbone	Depth	Res.	Pretrain	UCF101				HMDB51			
					R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
VCLR [29]	R2D	50	224	K400	70.6	80.1	86.3	90.7	35.2	58.4	68.8	79.8
RSPNet [6]	R3D	18	112	K400	41.1	59.4	68.4	77.8	-	-	-	-
MLRep [43]	R3D	18	112	K400	41.5	60.0	71.2	80.1	20.7	40.8	55.2	68.3
FAME [10]	R(2+1)D	10	112	K400	62.3	75.1	80.9	86.9	-	-	-	-
SpeedNet [3]	S3D-G	23	224	K400	13.0	28.1	37.5	49.5	-	-	-	-
HiCo(Ours)	R3D	18	112	UK400	**71.8**	**83.8**	88.5	92.8	**35.8**	59.7	71.1	81.2
HiCo(Ours)	R(2+1)D	10	112	UK400	69.1	84.4	**89.0**	**93.6**	35.2	58.8	70.3	**82.3**
HiCo(Ours)	S3D-G	23	112	UK400	62.5	76.4	82.9	89.4	35.5	**60.3**	**72.2**	82.1

Table 6. Nearest neighobor retrieval comparison on UCF101 and HMDB51. 'Grey fonts' refer different backbones with HiCo.

Method	Encoder	PT Data.	AUC	AR@100
VINCE [16]	R2D-50	K400	64.6%	73.2%
SeCo [68]	R2D-50	K400	65.2%	73.4%
VCLR [29]	R2D-50	K400	65.5%	73.8%
CL	S3D-G	UK400	63.0%	72.4%
HiCo	S3D-G	UK400	**67.1%**	**75.4%**

Table 7. Action localization on ActivityNet [12]. 'PT Data.' refers to pre-training dataset.

a slightly lower performance, but the gap is significantly closed when fully fine-tuning is employed. Second, when pre-training without freezing backbones, HiCo achieves better performances than previous approaches under similar settings. For example, using same input resolutions (16×112^2) and backbone (R(2+1)D), HiCo is 1.7% and 2.1% higher than FAME [10] on UCF101 and HMDB51, respectively. Note that ρBYOL [13] obtains excellent performance. The reason may be that it applies a different self-supervised learning method (BYOL), deeper network, and large resolution. When adopting the same SimCLR, HiCo can achieve comparable performance with ρSimCLR [13], using lower resolutions and a tinyer backbone. Third, compared to VCLR [29] trained on HACS using R3D-18, a notable improvement is observed with similar conditions on both datasets, with a gap of 9.8% and 6.9% on respective datasets. This demonstrates that HiCo is a more suitable framework for learning from untrimmed videos.

4.3. Evaluation on video retrieval task

We compute normalized cosine similarity with features extracted by UK400 pre-trained networks for video retrieval. Tab. 6 compares HiCo with other approaches with different top-k accuracies. HiCo exceeds the state-of-the-art method (*i.e.*, VCLR [29]) by 1.2% on UCF101 under R@1 with a lightweight R3D-18 network, which implies that features learned by HiCo is more generalized.

4.4. Evaluation on action localization task

We use the mainstream TAL method, *i.e.*, BMN [32], to evaluate the UK400 pre-trained features on ActivityNet [12]. As shown in Tab. 7, HiCo with a lightweight encoder significantly outperforms VCLR [29] by 1.6% in terms of AUC. Compared with standard contrastive learning, HiCo boosts the AUC by 4.1%. The main reason is that HiCo can preserve more high-level information for clips through topically consistent learning, which assists BMN in discriminating the actions and background. The results successfully demonstrate the transferability of HiCo pre-trained representations to different downstream tasks.

5. Discussions

Limitations. HiCo provides a simple framework for learning representations in untrimmed videos. Despite its effectiveness on existing public datasets, HiCo may fail when it encounters videos with various topics and more complex relations between different clips, such as movies or TV series. Therefore, for learning from unlabelled untrimmed videos, one should avoid using videos sampled from those sources.

Conclusion. In this paper, we propose HiCo, a novel self-supervised learning framework for learning powerful video representations from untrimmed videos. It exploits the hierarchical consistency existing in the long videos, *i.e.,* the visual consistency and the topical consistency. For visual consistency learning, HiCo employs the contrastive learning with constrained clip distances. For topical consistency learning, a topic prediction task is presented. Further, a gradual sampling strategy is proposed based on curriculum learning for both tasks, whose superiority is demonstrated theoretically and empirically. In general, we believe that untrimmed videos are not only easier to collect, but also provide more potential for learning more robust video representations. We hope the simple and effective framework of HiCo can encourage and inspire the researchers to be further devoted to this area.

Acknowledgement. This work is supported by the National Natural Science Foundation of China under grant 61871435, Fundamental Research Funds for the Central Universities no.2019kfyXKJC024, by the 111 Project on Computational Intelligence and Intelligent Control under Grant B18024, by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046) and by Alibaba Group through Alibaba Research Intern Program.

References

[1] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020. 2

[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Shot and scene detection via hierarchical clustering for re-using broadcast video. In *International Conference on Computer Analysis of Images and Patterns*, pages 801–811. Springer, 2015. 2

[3] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *CVPR*, pages 9922–9931, 2020. 2, 7, 8

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009. 4

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1, 5

[6] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. *arXiv preprint arXiv:2011.07949*, 2020. 7, 8

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 5

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[9] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic action and motion network. In *ICCV*, pages 6192–6201, 2019. 2

[10] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, and Jue Wang. Motion-aware self-supervised video representation learning via foreground-background merging. *arXiv preprint arXiv:2109.15130*, 2021. 2, 7, 8

[11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 2

[12] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 5, 8

[13] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 1, 2, 7, 8

[14] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, pages 3636–3645, 2017. 2

[15] Jialin Gao, Zhixiang Shi, Guanshuo Wang, Jiani Li, Yufeng Yuan, Shiming Ge, and Xi Zhou. Accurate temporal action proposal generation with relation-aware pyramid network. In *AAAI*, pages 10810–10817, 2020. 2

[16] 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. 8

[17] Michael Gygli. Ridiculously fast shot boundary detection with fully convolutional neural networks. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE, 2018. 2

[18] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 2

[19] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 7

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 5

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 2

[22] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 2

[23] Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. *arXiv preprint arXiv:2106.02342*, 2021. 7

[24] Ziyuan Huang, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Rong Jin, and Marcelo H Ang. Self-supervised motion learning from static images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1276–1285, 2021. 2

[25] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9970–9980, 2021. 2

[26] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 425–442. Springer, 2020. 2

[27] H Jhuang, H Garrote, E Poggio, T Serre, and T Hmdb. A large video database for human motion recognition. In *ICCV*, volume 4, page 6, 2011. 1, 2, 5

[28] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. Cycle-contrast for self-supervised video representation learning. *arXiv preprint arXiv:2010.14810*, 2020. 7

[29] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Soren Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 3195–3204, 2021. 2, 7, 8

[30] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 2

[31] Fu Li, Chuang Gan, Xiao Liu, Yunlong Bian, Xiang Long, Yandong Li, Zhichao Li, Jie Zhou, and Shilei Wen. Temporal modeling approaches for large-scale youtube-8m video understanding. *arXiv preprint arXiv:1707.04555*, 2017. 2

[32] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3889–3898, 2019. 2, 8

[33] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[35] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2203–2212, 2017. 2

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

[37] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 2

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

[40] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 2, 7

[41] Bo Pang, Gao Peng, Yizhuo Li, and Cewu Lu. Pgt: A progressive method for training models on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11389, 2021. 2

[42] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 2

[43] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7990–8001, 2021. 7, 8

[44] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 1, 2, 5, 7

[45] Zhiwu Qing, Ziyuan Huang, Shiwei Zhang, Mingqian Tang, Changxin Gao, Marcelo H Ang Jr, Rong Jin, and Nong Sang. Paramcrop: Parametric cubic cropping for video contrastive learning. *arXiv preprint arXiv:2108.10501*, 2021. 1, 2

[46] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 485–494, 2021. 2

[47] Hong Shao, Yang Qu, and Wencheng Cui. Shot boundary detection algorithm based on hsv histogram and hog feature. In *5th International Conference on Advanced Engineering Materials and Technology*, pages 951–957, 2015. 2

[48] Mike Zheng Shou, Stan W Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. *arXiv preprint arXiv:2101.10511*, 2021. 2

[49] 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. 1, 2, 5

[50] Tomáš Souček, Jaroslav Moravec, and Jakub Lokoč. Transnet: A deep network for fast detection of common shot transitions. *arXiv preprint arXiv:1906.03363*, 2019. 2

[51] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015. 2

[52] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020. 2

[53] Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. Fast video shot transition localization with deep structured models. In *Asian Conference on Computer Vision*, pages 577–592. Springer, 2018. 2

[54] Yongyi Tang, Xing Zhang, Lin Ma, Jingwen Wang, Shaoxiang Chen, and Yu-Gang Jiang. Non-local netvlad encoding for video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[55] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Selfsupervised video representation using pretext-contrastive learning. *arXiv preprint arXiv:2010.15464*, 2, 2020. 2

[56] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. 2020. 2

[57] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal

convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 5

[58] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 5

[59] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016. 2

[60] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4006–4015, 2019. 2

[61] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, pages 504–521. Springer, 2020. 2

[62] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 2

[63] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 2

[64] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 5

[65] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, pages 10334–10343, 2019. 2

[66] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020. 2

[67] Xitong Yang, Haoqi Fan, Lorenzo Torresani, Larry S Davis, and Heng Wang. Beyond short clips: End-to-end video-level learning with collaborative memories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7567–7576, 2021. 2

[68] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. *arXiv preprint arXiv:2008.00975*, 6(7), 2020. 8

[69] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, pages 6548–6557, 2020. 2

[70] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 2

[71] Lin Zhang, Qi She, Zhengyang Shen, and Changhu Wang. How incomplete is contrastive learning? an inter-intra variant dual representation method for self-supervised video recognition. *arXiv preprint arXiv:2107.01194*, 2021. 7

[72] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016. 2

[73] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. Hacs: Human action clips and segments dataset for recognition and temporal localization. *arXiv preprint arXiv:1712.09374*, 2019. 5

[74] Dimitri Zhukov, Jean-Baptiste Alayrac, Ivan Laptev, and Josef Sivic. Learning actionness via long-range temporal order verification. In *European Conference on Computer Vision*, pages 470–487. Springer, 2020. 2