Static and Dynamic Concepts for Self-supervised Video Representation Learning

Rui Qian[1], Shuangrui Ding[2], Xian Liu[1], and Dahua Lin[1,3]*

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Shanghai Artificial Intelligence Laboratory, Shanghai, China
{qr021,lx021,dhlin}@ie.cuhk.edu.hk dsr1212@sjtu.edu.cn

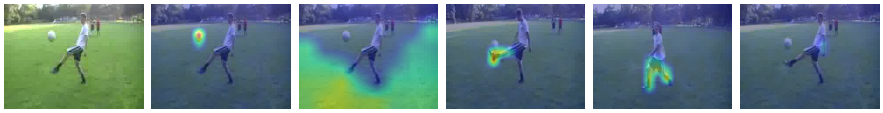
Abstract. In this paper, we propose a novel learning scheme for self-supervised video representation learning. Motivated by how humans understand videos, we propose to first learn general visual concepts then attend to discriminative local areas for video understanding. Specifically, we utilize static frame and frame difference to help decouple static and dynamic concepts, and respectively align the concept distributions in latent space. We add diversity and fidelity regularizations to guarantee that we learn a compact set of meaningful concepts. Then we employ a cross-attention mechanism to aggregate detailed local features of different concepts, and filter out redundant concepts with low activations to perform local concept contrast. Extensive experiments demonstrate that our method distills meaningful static and dynamic concepts to guide video understanding, and obtains state-of-the-art results on UCF-101, HMDB-51, and Diving-48.

Keywords: Video Representation · Visual Concepts · Local Contrast

1 Introduction

Self-supervised representation learning has been an exciting problem in computer vision, which aims to encode robust representations that can be transferred to various downstream tasks without human labeling. A prevalent strategy is to design pretext tasks and acquire pseudo labels as self-supervision [7,24] or employ contrastive learning to discriminate instances [12,29,8]. However, this learning scheme is inconsistent with how humans learn from the world. To be specific, instead of solely learning from labels or contrasting global features, humans can typically conclude some general basic concepts from detailed observations, then make predictions based on these concepts [6,62,38]. For example, we identify an airplane through its wings and rudder; and recognize the action of playing soccer through the ball as well as running and kicking movement as in Fig 1. To this end, it would be promising to automatically formulate transferable concepts to guide detailed local feature perception and improve the representations.

* Corresponding author. Email: dhlin@ie.cuhk.edu.hk



(a) Soccer Juggling

(b) Basketball

Fig. 1. Visualization of visual concept attention maps. Each column corresponds to the same concept, the former two columns describe static concepts and the latter three present dynamic concepts. The same visual concept highlights similar visual patterns, e.g., spherical objects, grass land, foot movement, leg movement, arm movement.

There have been some works exploring learning interpretable visual concepts for particular tasks [38,6,78,13]. But in unsupervised video representation learning, how to formulate meaningful visual concepts and efficiently leverage local cues remains unsolved. The difficulty lies in two aspects: Videos contain more redundancy on temporal dimension. Besides, we lack fine-grained supervision on the potential visual concepts. Most of the recent state-of-the-art works on video representation learning inherit contrastive learning framework [58,28,22], which projects the global pooled feature vectors into a latent space and performs instance discrimination. Compared with the aforementioned human perception, this formulation explicitly contrasts high-level global feature vectors but has difficulty dealing with detailed local features. Some works propose region-based local feature contrast but could result in high redundancy [73,83]. In order to effectively utilize the detailed local features, we propose a novel learning strategy for self-supervised video representation learning. We aggregate local features that present similar concepts, and then perform the concept-level alignment.

Concretely, we propose to form a latent space consisting of the learned visual concepts, and leverage the latent concept distributions as self-supervision to jointly optimize feature representations and concept descriptions. However, since the feature attributes are highly entangled in the high-level representation, it is nontrivial to directly obtain general concepts without annotations. To solve this, we divide the learning concepts into two general divisions, i.e., static scenes and dynamic motions. Those two concepts are proved to be complementary but orthogonal for video representation learning [32]. Static scenes focus on background cues while dynamic motions lay more emphasis on object's movement. In practice, we use the simple static frame and frame difference to naturally decouple these two aspects and ameliorate the entanglement of high-level feature. Further, we define the projection head as a cosine classifier to generate concept latent codes, with each class corresponding to a potential static or dynamic local concept. We respectively align the static (dynamic) concept latent codes between

original video and static frame (frame difference), and encourage sparsity in the latent space to guarantee diversity of learned concepts. Besides, to make the projection head preserve necessarily relevant information and reduce redundancy, we regard the latent codes as information bottleneck, where they are expected to reconstruct the initial feature vectors. Thus, we apply a light-weight MLP to achieve the fidelity regularization. By doing so, we establish a concept-based latent space consisting of general static and dynamic visual concepts.

With these learned concept prototypes, we attend to local concepts in each spatio-temporal area to improve the detailed local feature modeling. Specifically, we use cross-attention to aggregate local features, then output a set of features belonging to different concepts like Fig. 1. By referring to the concept latent code, we select a series of visual concepts with high activations as valid ones and filter out the redundant feature pairs. Contrastive loss is applied to these valid pairs for fine-grained alignment. In this way, we seamlessly integrate general concept learning with detailed local feature perception to enhance video representations.

To sum up, our contributions are: (1) We propose a novel self-supervised video representation learning scheme, where we formulate general concepts to guide concept-level detailed local feature alignment. (2) We employ cross-attention to aggregate detailed features of different concepts, and filter out redundant local features by concept latent codes. In this way, we achieve efficient local concept contrast. (3) We achieve state-of-the-art results on downstream action recognition and video retrieval across UCF-101, HMDB-51 and Diving-48 datasets.

2 Related Work

Self-supervised Learning. Self-supervised learning aims to make full use of large-scale unlabelled data without resorting to human annotations. Some works design pretext tasks, e.g., image rotation [24], colorization [37], clustering [7,60], to obtain pseudo labels and guide representation learning. Another line of works introduce contrastive learning to build robust feature representations [75,20,52]. They employ noise contrastive estimation [25] to compare feature representations and discriminate different instances [65,12,29]. Technically, these methods rely on nonlinear projection head to project the extracted features into a latent space for contrastive loss computation to reduce information loss. However, without explicit constraint on the projection head, what information is preserved and contrasted in the latent space is unclear, and the learning process is of low interpretability. More recently, [21] employs whitening to analyze the latent feature space. [8] assigns features to prototype vectors and contrasts cluster assignments in the latent space. In contrast, in this work, we enforce the projection heads to learn potential visual concepts and formulate an interpretable latent space, where we contrast the concept distributions to guide general representation learning.

Video Representation Learning. Representation learning in video domain requires the model to capture crucial spatio-temporal relationships in video sequences. Early works employ the temporal transformation [51,79,81,5,35,11,80], spatio-temporal jigsaw [36,69], temporal cycle-consistency [33,42,72], future pre-

diction [4,49] as pretext tasks. Later, [58,70,22,55,46,44,30,45] expand contrastive learning framework to video and audio-video domain. Further, [34,39,31,17,18] utilize the internal temporal structure to generate richer positive samples. [28,41,2,50,54] contrast temporally aligned multi-modal inputs to learn complementary information. These works explicitly contrast the global representations of video clips, but pay little attention to detailed local features. To this end, [26,27] propose to predict dense feature maps in future timestamps. [59,3,15,56] contrast short and long clips on each timestamp to attend to fine-grained temporal features, but still fail to utilize detailed spatial cues. [83,10] rely on bounding boxes or segmentation masks to align semantically related local areas. While in our work, we use simple static frame and frame difference to distill static and dynamic visual concepts, based on which we aggregate relevant information from each spatio-temporal area to enhance detailed content modeling.

Concept Learning. Recently, there have emerged a line of works that learn human-specified visual concepts to solve downstream visual tasks [38,13,6,47,16]. They design concept bottlenecks models to first predict concepts then use these concepts to make final predictions. Comparing to end-to-end deep models, concept bottleneck models are more interpretable but require extra concept annotations. To tackle this problem, [61,1] develop various regularizations to constrain the concept bottleneck and obtain potential concepts. [78] points out that one-hot category labels are not optimal concept descriptions, and devises an exploration-experience loss to alternatively update feature representation and concept description. To our best knowledge, we are the first to integrate concept learning into self-supervised video representation learning. We utilize static and dynamic visual concepts to learn both general and detailed video representations.

3 Method

Our framework is shown in Fig. 2. For simplicity, we show detailed procedures for video clip v, while static frame s and d are processed similarly. Specifically, we first propose decoupled concept alignment (Sec. 3.1) with regularizations (Sec. 3.2) to jointly optimize the extracted features and concept descriptions. Then referring to learned concepts, we employ cross-attention to aggregate detailed local features of different concepts, filter out redundant concepts with low activations and perform concept-level alignment (Sec. 3.3).

3.1 Decoupled Concept Learning

Videos typically possess two complementary concepts, static concepts that indicate background scene attributes, and dynamic concepts that reveal human or object movements. Given a video sequence v, since various visual concepts are highly entangled, it is nontrivial to directly learn meaningful visual concepts without resorting to human annotations. But it is practical to decouple the static and dynamic information in the input stage, i.e., we randomly select a static frame s and calculate frame difference d to respectively carry static and

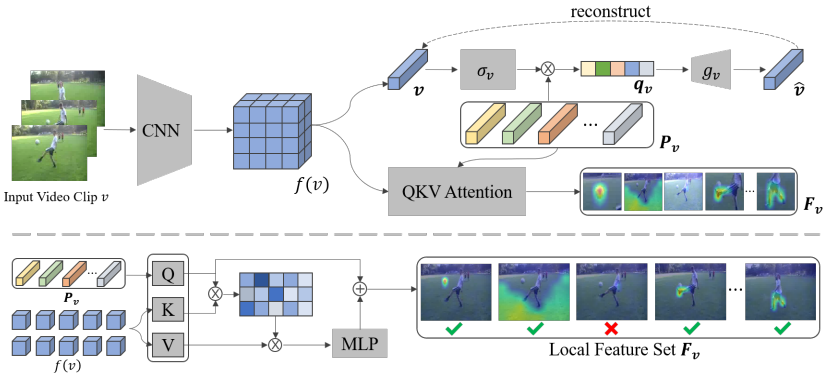


Fig. 2. Overview of the framework. We take the original input video clip v for illustration. In the upper branch, we calculate cosine similarity between concept prototypes $\boldsymbol{P_v}$ and transformed video feature $\sigma_v(\boldsymbol{v})$ as concept latent code $\boldsymbol{q_v}$, which is then passed through simple MLP to reconstruct the original feature vector. In the lower branch, we use QKV attention with residue to aggregate local features of different concepts and refer to $\boldsymbol{q_v}$ to avoid redundant local concept contrast.

dynamic attributes. Then, an intuitive idea is to learn potential static concepts from v and s, extract dynamic concepts from v and d, and respectively perform static and dynamic concept alignment.

Concept Prototypes. To formulate the latent concept space, we propose to learn several prototypes, each corresponding to a static or dynamic concept. Specifically, we define three sets of prototypes respectively for s, d, v as:

$$\boldsymbol{P_s} \in \mathbb{R}^{K_s \times C}, \quad \boldsymbol{P_d} \in \mathbb{R}^{K_d \times C}, \quad \boldsymbol{P_v} \in \mathbb{R}^{(K_s+K_d) \times C}, \tag{1}$$

where C denotes channel dimension, K_s is the number of static concepts, K_d is the number of dynamic concepts. We use these concept prototypes to generate latent concept activation codes and retrieve relevant local features in later stage.

Concept Codes. Following [12,8], we use a projection head to project the features into a latent space and generate the concept latent codes. Mathematically, we denote the feature extractor as f, and employ global average pooling to obtain three feature vectors[4]:

$$\boldsymbol{s} = GAP(f(s)), \quad \boldsymbol{d} = GAP(f(d)), \quad \boldsymbol{v} = GAP(f(v)), \tag{2}$$

each is of the same dimension $\mathbb{R}^C$. Then, we pass these feature vectors through projection heads to calculate concept codes. For illustration, we take the concept code $\boldsymbol{q_s}$ for static frame as an example. We first input $\boldsymbol{s}$ into a transformation σ_s, which is in default identity mapping but can be replaced with other shallow layers like MLP. Then we calculate the cosine similarity between the output

[4] For simplicity, we use the same symbol to denote the backbone for s, d, v.

vector and each prototype to form $\boldsymbol{q_s}$:

$$\boldsymbol{q_s}^{(k)} = \frac{\boldsymbol{P_s}^{(k)} \sigma_s(\boldsymbol{s})^T}{||\boldsymbol{P_s}^{(k)}||_2 ||\sigma_s(\boldsymbol{s})||_2}, \quad \boldsymbol{q_s} \in \mathbb{R}^{K_s}, \tag{3}$$

where the superscript (k) indicates k-th channel. Similarly, we obtain concept codes $\boldsymbol{q_d} \in \mathbb{R}^{K_d}$ and $\boldsymbol{q_v} \in \mathbb{R}^{K_s+K_d}$ in the same manner.

Concept Alignment. Since the video and static frame share the same static attributes, while video and frame difference have the same dynamic attributes, we propose to respectively align the static and dynamic concepts through Eq. 4.

$$\begin{aligned} \mathcal{L}_{aln} = &- \sum_{k=1}^{K_s} \left(\overline{\boldsymbol{q_s}}^{(k)} \log \frac{\exp(\boldsymbol{q_v^s}^{(k)}/\tau)}{\sum_{k'} \exp(\boldsymbol{q_v^s}^{(k')}/\tau)} + \overline{\boldsymbol{q_v^s}}^{(k)} \log \frac{\exp(\boldsymbol{q_s}^{(k)}/\tau)}{\sum_{k'} \exp(\boldsymbol{q_s}^{(k')}/\tau)} \right) \\ &- \sum_{k=1}^{K_d} \left(\overline{\boldsymbol{q_d}}^{(k)} \log \frac{\exp(\boldsymbol{q_v^d}^{(k)}/\tau)}{\sum_{k'} \exp(\boldsymbol{q_v^d}^{(k')}/\tau)} + \overline{\boldsymbol{q_v^d}}^{(k)} \log \frac{\exp(\boldsymbol{q_d}^{(k)}/\tau)}{\sum_{k'} \exp(\boldsymbol{q_d}^{(k')}/\tau)} \right), \end{aligned} \tag{4}$$

For simplicity, we divide $\boldsymbol{q_v}$ into two parts, the former K_s channels as $\boldsymbol{q_v^s}$ indicating static concepts, and the latter K_d channels as $\boldsymbol{q_v^d}$ for dynamic concepts. Similar to SWAV [8], we assume the concepts follow a uniform distribution over the whole dataset, and use Sinkhorn-Knopp algorithm [14] to generate the soft code $\overline{\boldsymbol{q}}$. Then we calculate the cross-entropy between $\overline{\boldsymbol{q}}$ and the latent concept distribution by taking softmax with temperature τ on $\boldsymbol{q}$. By minimizing $\mathcal{L}_{aln}$, we respectively align static and dynamic concept distributions, and jointly optimize feature representations and concept descriptions from large-scale video data.

3.2 Concept Bottleneck Constraint

However, the decoupled concept alignment objective alone cannot guarantee that each of the learned prototype corresponds to a meaningful concept. Motivated by [1], the general concepts should possess fidelity and diversity. That is, the concepts should preserve much relevant information from the inputs, and the inputs can be described by a few concepts. To this end, we devise two constraints on the concept latent codes as follows.

The first constraint is the sparsity regularization term as Eq. 5 to enforce diversity of learned concepts. We employ $\mathcal{L}_1$ norm regularization to encourage sparsity of concept latent codes, so that each input activates only a few concepts.

$$\mathcal{L}_{div} = \|\boldsymbol{q_s}\|_1 + \|\boldsymbol{q_d}\|_1 + \|\boldsymbol{q_v}\|_1 . \tag{5}$$

The second constraint is a reconstruction loss as Eq. 6 to ensure fidelity and reduce redundancy. We borrow the idea from autoencoder to reconstruct the feature vectors. Since the channel dimension of concept code is smaller than the

feature vector, we regard $\boldsymbol{q}$ as information bottleneck and pass them through two-layer MLP g for reconstruction. We use $\mathcal{L}_2$ loss for optimization, and stop gradient on the original features. In this way, the concept prototypes cover a wide range of important information with low redundancy.

$$\mathcal{L}_{fid} = \|g_s(\boldsymbol{q_s}) - \boldsymbol{s}\|_2^2 + \|g_d(\boldsymbol{q_d}) - \boldsymbol{d}\|_2^2 + \|g_v(\boldsymbol{q_v}) - \boldsymbol{v}\|_2^2 . \tag{6}$$

Relation to SWAV. Our concept code formulation is similar to SWAV [8], both using cosine similarity between feature vectors and prototypes. But the motivations and technical designs are different. In terms of the motivation, SWAV is essentially over-clustering and the prototypes are cluster centroids, the number of which is set as 3,000 in default, much greater than semantic categories. While in our method, the prototypes project the feature vectors into the low dimensional space, which interprets the concept activations instead of the instance discrimination. Through regularizations and activation alignment, our prototypes are an ordered set of interpretable concepts each presenting a visual attribute. In terms of technical design, our method only conducts spatio-temporal cropping due to multiple modalities while SWAV requires stronger augmentation to make the pretraining task harder and improve the representation quality.

3.3 Local Concept Contrast

The global concept code alignment serves as an effective supervision to learn spatio-temporal characters in videos, but does not make use of detailed local features which are crucial for video understanding. Some existing works in image domain first match corresponding local areas then make contrast [73,77], but they have difficulty expanding to videos because of the redundancy on time dimension. [83] employs bounding boxes for region-based contrast between video clips, but requires prior to filter redundant background areas. In order to better utilize the detailed local contents, we need to generate a compact set of local features with low redundancy. Therefore, we propose to leverage the learned prototypes to retrieve detailed local features that are relevant to particular concepts, and output an ordered set of local features for effective contrast.

Local Feature Attention. Motivated by the success of attention mechanism in local feature aggregation [67,19,71,23], we employ widely used cross-attention mechanism to retrieve detailed local features that are relevant to specific visual concepts. As illustrated in Fig. 2, we linearly project the concept prototypes as query tokens, and project the feature maps to formulate key and value tokens. Then QKV attention with residue is applied to aggregate local features related to the query. We still use the local features on static frame as an example:

$$\boldsymbol{F_s} = QKV(\boldsymbol{P_s}, f(\boldsymbol{s}), f(\boldsymbol{s})), \quad \boldsymbol{F_s} \in \mathbb{R}^{K_s \times C}. \tag{7}$$

We obtain $\boldsymbol{F_d} \in \mathbb{R}^{K_d \times C}$ and $\boldsymbol{F_v} \in \mathbb{R}^{(K_s+K_d) \times C}$ in the same manner. Similar to the separation on $\boldsymbol{q_v}$, we also divide $\boldsymbol{F_v}$ into $\boldsymbol{F_v^s} \in \mathbb{R}^{K_s \times C}$ and $\boldsymbol{F_v^d} \in \mathbb{R}^{K_d \times C}$.

Since each prototype corresponds to a potential static or dynamic concept, each generated attention map highlights local areas that contain particular concepts as shown in Fig. 1, where each column belongs to the same concept. Therefore, it is intuitive to apply contrastive loss on the aggregated features of the matching concepts to further enhance detailed local representations.

Local Feature Contrast. Recall that each input is representable with a few concepts, we need to first filter out a set of valid concepts that exist in the input sample. To do this, we resort to the previously obtained concept latent codes $\boldsymbol{q}$, which figure out which concepts are activated in each training sample. Mathematically, we take local features of static concepts for illustration. Given concept latent codes $\boldsymbol{q_s}, \boldsymbol{q_v^s}$ and local features $\boldsymbol{F_s}, \boldsymbol{F_v^s}$, we select top-$K$ indexes of each latent code and take the intersection as the valid static concept indexes:

$$\boldsymbol{idx_s} = \text{top-k}(\boldsymbol{q_s}, K) \cap \text{top-k}(\boldsymbol{q_v^s}, K). \tag{8}$$

The valid local feature pairs are denoted as $\{(\boldsymbol{F_s}^{(k)}, \boldsymbol{F_v^s}^{(k)}) | k \in \boldsymbol{idx_s}\}$, with the superscript (k) indicating local feature of k-th concept.

These local features of the same static (dynamic) concept from the same video are expected to represent exactly the same appearances (movements), thus should be aligned. To this end, we apply contrastive margin loss in Eq. 9 to contrast the local features of valid concept indexes. To be specific, we employ the valid local feature pair from the same video as positive samples, and use local features of corresponding concept from other videos in the mini-batch to form negative samples. We minimize the $\mathcal{L}_2$ distance between positive feature pairs, and push the distance between negative pairs to a large margin:

$$l(\boldsymbol{F_s}, \boldsymbol{F_v^s}) = \sum_{k \in \boldsymbol{idx_s}} \left[\left\| \boldsymbol{F_s}^{(k)} - \boldsymbol{F_v^s}^{(k)} \right\|_2^2 + \sum_{\widetilde{\boldsymbol{F}} \in \mathcal{N}} \max\left(\lambda - \left\| \boldsymbol{F_s}^{(k)} - \widetilde{\boldsymbol{F}}_{\boldsymbol{v}}^{\boldsymbol{s}(k)} \right\|_2, 0 \right)^2 \right], \tag{9}$$

where λ is the margin hyper-parameter, and $\mathcal{N}$ is the set of negative samples in the mini-batch. We use similar techniques to process local features of dynamic concepts, and the final local concept contrast learning objective is formulated as

$$\mathcal{L}_{loc} = l(\boldsymbol{F_s}, \boldsymbol{F_v^s}) + l(\boldsymbol{F_v^s}, \boldsymbol{F_s}) + l(\boldsymbol{F_d}, \boldsymbol{F_v^d}) + l(\boldsymbol{F_v^d}, \boldsymbol{F_d}). \tag{10}$$

By minimizing $\mathcal{L}_{loc}$, we build a concept-level self-supervision to make use of detailed local features and improve video representations. Comparing to previous methods using similar techniques to contrast local features [83,74], our method does not rely on prior or complex post-processing to filter out redundant feature pairs. The integration of general concept learning and detailed local feature contrast leads to higher learning efficiency and more comprehensive representations.

Overall Learning Objective. The overall training objective can be written as

$$\mathcal{L} = \mathcal{L}_{aln} + \alpha \mathcal{L}_{loc} + \beta \mathcal{L}_{fid} + \gamma \mathcal{L}_{div}, \tag{11}$$

where the balancing hyper-parameters are respectively set to $\alpha = \beta = 1, \gamma = 0.01$ in default. Since the formulation of $\mathcal{L}_{loc}$ relies on the concept codes to filter out valid pairs, in the first few epochs (5 epochs in default), we do not include $\mathcal{L}_{loc}$ to prevent random selection and stabilize training.

4 Experiment

4.1 Dataset

We use 4 popular video datasets, Kinetics-400 [9], UCF-101 [63], HMDB-51 [40] and Diving-48 [43]. **Kinetics-400** [9] is a widely used benchmark for self-supervised video representation learning, with 240K video clips covering 400 human action classes. **UCF-101** [63] covers 101 action categories and more than 13K annotated clips. **HMDB-51** [40] contains around 7k clips covering 51 action classes. **Diving-48** [43] contains 48 different diving actions. Different action classes in Diving-48 mainly vary in motion patterns and the backgrounds are quite similar.

4.2 Implementation Details

We choose R(2+1)D-18 [66] with 14.4M parameters, and S3D [76] as the video encoder. We empirically find that using separate networks or sharing the same network to extract RGB/static frame/frame difference features leads to similar performance. But using shared backbone results in higher learning efficiency, so we use the same backbone for all in default. Given a video clip, we randomly select a frame and repeat 16 times on the temporal axis to construct static frame input, and use the difference between adjacent frames to form the frame difference input. The resolution of each input sequence is $16 \times 112 \times 112$ if not specially motioned. We pretrain the model for 200 epochs on UCF-101 or 100 epochs on Kinetics-400. We adopt SGD optimizer with the initial learning rate of 10^{-2} and weight decay of 10^{-4}. We set the number of static or dynamic concepts to $K_s = K_d = 50$, and the ratio of valid local concepts to 10%, $K = 5$ in default.

4.3 Evaluation on Downstream Tasks

Action Recognition. We first present action recognition in Table 1. We report *linear probe* and *finetune* Top-1 accuracy. For fair comparison, we exclude the works with different evaluation settings and much deeper backbone [58,22] or rely on audio and text [59,50]. The † means jointly utilizing RGB and optical flow for pretraining, and the final performance is tested with RGB only.

In *linear probe* settings, our method achieves state-of-the art results on both two datasets. It is worth noting that our UCF-101 pretrained model even outperforms most RGB-based methods pretrained on Kinetics-400, which indicates the high data efficiency of our learning framework. Regarding to comparison with CoCLR [28] pretrained with RGB and Flow, we reach higher accuracy with fewer frames in each clip. It indicates that simple frame difference could replace computationally expensive optical flow to improve dynamic attribute learning.

Method	Backbone	Pretrain Dataset	Frames	Res.	Freeze	UCF-101	HMDB-51
CBT [64]	S3D	Kinetics-600	16	112	✓	54.0	29.5
RSPNet [11]	R3D	Kinetics-400	16	112	✓	61.8	42.8
MLRep [57]	R3D	Kinetics-400	16	112	✓	63.2	33.4
CoCLR† [28]	S3D	Kinetics-400	32	128	✓	74.5	46.1
Ours	R(2+1)D	UCF-101	16	112	✓	67.4	40.7
Ours	R(2+1)D	Kinetics-400	16	112	✓	72.1	45.9
Ours	S3D	Kinetics-400	16	128	✓	75.1	47.4
TempTrans [35]	R(2+1)D	UCF-101	16	112	✗	81.6	46.4
LSFD [3]	R3D	UCF-101	32	112	✗	77.2	53.7
STS† [68]	R(2+1)D	UCF-101	16	112	✗	77.8	40.7
CoCLR† [28]	S3D	UCF-101	32	128	✗	81.4	52.1
Ours	R(2+1)D	UCF-101	16	112	✗	82.1	49.7
Ours	S3D	UCF-101	32	128	✗	83.7	53.8
ASCNet [31]	R3D	Kinetics-400	16	112	✗	80.5	52.3
Pace [70]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
VideoMoCo [53]	R(2+1)D	Kinetics-400	32	112	✗	78.7	49.2
RSPNet [11]	R(2+1)D	Kinetics-400	16	112	✗	81.1	44.6
TCLR [15]	R(2+1)D	Kinetics-400	16	112	✗	84.3	54.2
TimeEq [34]	S3D-G	Kinetics-400	32	128	✗	86.9	63.5
STS† [68]	S3D-G	Kinetics-400	64	224	✗	89.0	62.0
CoCLR† [28]	S3D	Kinetics-400	32	128	✗	87.9	54.6
Ours	R(2+1)D	Kinetics-400	16	112	✗	86.1	54.8
Ours	S3D	Kinetics-400	16	128	✗	88.3	56.4

Table 1. Results on action recognition downstream task. We present the backbone encoder, pretrain dataset, spatio-temporal resolution of each method. Freeze (tick) indicates *linear probe*, and no freeze (cross) denotes *end-to-end finetune*.

In *finetune*, ours also achieves the best results among RGB-only methods, and is comparable with RGB-Flow two-stream models. Among these method, [35,34,11,31] carefully design temporal transformations to enhance temporal perception in videos, [3,15] employ short and long clips to attend to fine-grained temporal features, [28,68] utilize complementary information between RGB and Flow to enhance video representations. While our method proposes to formulate general static and dynamic concepts to guide detailed local feature perception, the performance demonstrates the effectiveness of our new learning scheme.

Video Retrieval. We show the performance on video retrieval with R@k in Table 2. All models are pretrained on UCF-101 with resolution 112×112 for fair comparison. Generally, our method achieves superior results over both RGB-only and RGB-Flow two-stream methods, especially when k is small. It indicates that our method encodes desired characteristics into a more compact manifold.

4.4 Concept Analysis

Intuitively, actions can be represented by some general concepts, and the detailed feature description of these concepts help to discriminate similar action

Method	Backbone	UCF-101				HMDB-51			
		R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
VCP [48]	R3D	18.6	33.6	42.5	53.3	7.6	24.4	36.3	53.6
MLRep [57]	R3D	39.6	57.6	69.2	78.0	18.8	39.2	51.0	63.7
VCLR [39]	R2D-50	46.8	61.8	70.4	79.0	17.6	38.6	51.1	67.6
PRP [82]	R(2+1)D	20.3	34.0	41.9	51.7	8.2	25.3	36.2	51.0
STS† [68]	R(2+1)D	38.1	58.9	68.9	77.2	16.4	36.9	50.5	65.4
CoCLR† [28]	S3D	53.3	69.4	76.6	82.0	23.3	43.2	53.5	65.5
Ours	R(2+1)D	55.6	70.1	77.4	83.1	24.4	45.1	54.5	66.4

Table 2. Results on video retrieval downstream task. We report R@k (k=1,5,10,20), † means pretrained with RGB and optical flow.

classes. To this end, in this section, we reveal how the learned static and dynamic concepts influence downstream action recognition.

Feature	$\boldsymbol{v}$	$\boldsymbol{q_v}$	$\boldsymbol{q_v^s}$	$\boldsymbol{q_v^d}$	$\boldsymbol{F_v}$	$\boldsymbol{F_v^s}$	$\boldsymbol{F_v^d}$
UCF-101	72.1	66.3	61.4	62.6	72.7	68.3	69.8
HMDB-51	45.9	43.8	42.9	40.1	46.3	45.7	44.2
Diving-48	73.4	59.4	26.7	64.8	72.5	31.1	74.1

Table 3. Results of static and dynamic concept analysis. The models in first two rows and the third row are respectively pretrained on Kinetics-400 and Diving-48.

Decoupled Concept for Action Recognition. We first quantitatively analyze the static and dynamic concepts and their relevant local features on action recognition. We adopt different outputs from our learning framework and pass them through linear classifier to do action classification on UCF-101, HMDB-51 and Diving-48. Specifically, in default evaluation settings, we use the global average pooled $\boldsymbol{v}$ as input to the classifier. We also compare using the concept latent codes or the local feature set for recognition. Note that when using the local feature set, e.g., $\boldsymbol{F_v}$, we first filter out Top-10% concepts from $\boldsymbol{q_v}$, then average the corresponding local features for classification. From Table 3, we have several observations. First, using concept latent code for classification leads to performance drop, while the local feature set slightly improves performance. This is because we learn limited number of general concepts and could lose detailed information. While the local feature set effectively aggregates detailed information and drops redundant features, which helps to improve action recognition. Second, on UCF-101 and HMDB-51, static and dynamic concepts are almost of equal significance, and jointly utilizing static and dynamic concepts leads to best performance. Third, the dynamic concepts dominate action recognition on Diving-48, and the static concepts are nearly useless as expected. This is because different diving classes share the same background scene and only differ in motions, the static concepts could disturb motion pattern discrimination.

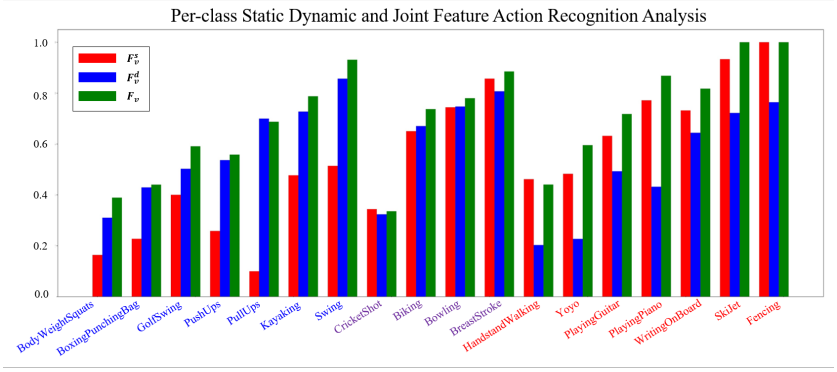
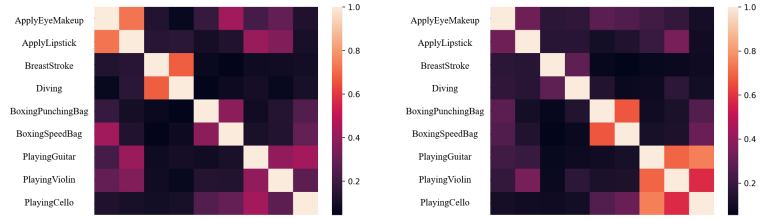


Fig. 3. Per-class action recognition accuracy analysis. We compare the performance of using static, dynamic and joint concept related local feature set, namely $\boldsymbol{F_v^s}$, $\boldsymbol{F_v^d}$, $\boldsymbol{F_v}$.

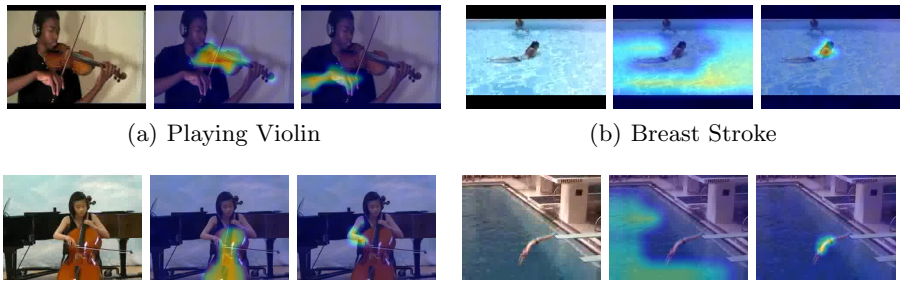


(a) Static concept code (b) Dynamic concept code

Fig. 4. Decoupled concept code similarity. We respectively average static and dynamic concept latent codes, $\boldsymbol{q_v^s}$ and $\boldsymbol{q_v^d}$, within each category, then calculate cosine similarity.

To further analyze the impact of the decoupled concepts on specific action categories, we select some typical classes from UCF-101 and visualize the per-class accuracy under different settings in Fig. 3. Among the selected action categories, the blue ones are highly dominated by motions, the red ones may have ambiguous motion patterns but can be easily recognized by appearance, and for the purple ones, both static appearance and dynamic motion are discriminative. The per-class accuracy with different feature input is in line with our expectations, which indicates the decoupled concepts respectively reveal static and dynamic attributes. Besides, we analyze the inter-class similarity in static and dynamic concept latent space. In Fig. 4, we visualize the similarity between different actions. Intuitively, some actions share similar background but with different motions, e.g., breaststroke and diving, while some possess similar movement but diverse appearances, e.g., playing different instruments. As expected, the former ones have higher inter-class similarity in static concept space while the latter ones are more similar regarding to dynamic concepts.

Visualization Results. For each clip, we respectively select a static and a dynamic concept with highest activation in latent space, and visualize the attention maps. Generally, the selected static concept attends to foreground objects or representative scene components, while the selected dynamic concepts highlights discriminative motions. Comparing Fig. 5(a) and Fig. 5(c), they share similar dynamic attributes, i.e., almost synchronized forearm movements, but can be discriminated by static objects. Regarding to Fig. 5(b) and Fig. 5(d), they happen in similar pools, but the dynamic concept helps to figure out distinct motion patterns. It reveals that the we learn meaningful static and dynamic concepts that focus on different aspects, these two jointly facilitate video understanding.



(a) Playing Violin (b) Breast Stroke

(c) Playing Cello (d) Diving

Fig. 5. Visualization of static and dynamic concept attention maps. Each subfigure left to right is: original frame, static concept attention map, dynamic concept attention map.

4.5 Ablation Study

We perform ablation studies on the loss function designs and crucial hyper-parameters. More ablative experiments please refer to Supplementary Material.

$\mathcal{L}_{aln}$	$\mathcal{L}_{fid}$	$\mathcal{L}_{div}$	$\mathcal{L}_{loc}$	UCF-101		HMDB-51	
				Linear	Finetune	Linear	Finetune
✓				61.4	76.3	40.3	44.7
✓	✓	✓		68.1	80.1	43.2	47.9
✓			✓	67.4	78.9	43.3	46.4
✓	✓	✓	✓	72.1	82.1	45.9	49.7

Table 4. Ablation study on loss functions. We pretrain on Kinetics-400.

Overall Framework. We first validate the effectiveness of the loss functions in Table 4. The model is pretrained with default concept numbers, $K_s = K_d = 50$,

and the decoupled concept alignment objective $\mathcal{L}_{aln}$ serves as the baseline. We can observe that the two regularizations $\mathcal{L}_{fid}$ and $\mathcal{L}_{div}$ significantly improve the performance. This is because these two terms reduce redundancy in the learned concept prototypes and contribute to more compact and diverse concept formulation, which effectively guides representation learning. And regarding to $\mathcal{L}_{loc}$, it also brings significant improvement since this objective explicitly contrasts local features of valid concepts and facilitates detailed local feature perception.

K_s	K_d	UCF-101		HMDB-51	
		w/ $\mathcal{L}_{loc}$	w/o $\mathcal{L}_{loc}$	w/ $\mathcal{L}_{loc}$	w/o $\mathcal{L}_{loc}$
25	25	70.3	61.2	43.0	39.4
25	50	71.7	66.3	44.1	40.8
50	25	71.3	65.2	44.8	42.4
50	50	72.1	68.1	45.9	43.2
100	100	72.3	68.8	45.8	44.3
200	200	72.3	69.4	45.6	44.1

Table 5. Ablation study on concept numbers. We report linear probe accuracy.

Number of Concepts. We also explore the impact of different concept numbers in Table 5. With the help of $\mathcal{L}_{loc}$, the performance slightly improves when K_s and K_d increases, and maintains stable in range of 50 to 200. While without $\mathcal{L}_{loc}$, the performance dramatically drops when the concept numbers become small. Because when K_s and K_d are small, the latent space captures general concepts but loses detailed information to discriminate similar actions. But when combined with $\mathcal{L}_{loc}$, the model adaptively attends to detailed local features with desired concepts, which makes up for the information loss to a large extent.

5 Conclusion

In this paper, we propose to learn general static and dynamic visual concepts to guide self-supervised video representation learning. We design decoupled concept alignment objective with regularizations to jointly optimize feature representations and concept distributions. Then we refer to the learned concepts to aggregate detailed local features corresponding to different concepts. We utilize the concept latent code to filter out redundant concepts with low activations, and perform concept-level local feature contrast for detailed video understanding. We achieve state-of-the-art results on UCF-101, HMDB-51 and Diving-48. The ablation studies demonstrate that the integration of general concept learning and detailed local feature contrast improves video representation learning.

Acknowledgement. This work is supported by GRF 14205719, TRS T41-603/20-R, Centre for Perceptual and Interactive Intelligence, and CUHK Interdisciplinary AI Research Institute.

References

1. Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems **31** (2018)
2. Asano, Y.M., Patrick, M., Rupprecht, C., Vedaldi, A.: Labelling unlabelled videos from scratch with multi-modal self-supervision. arXiv preprint arXiv:2006.13662 (2020)
3. Behrmann, N., Fayyaz, M., Gall, J., Noroozi, M.: Long short view feature decomposition via contrastive video representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9244–9253 (2021)
4. Behrmann, N., Gall, J., Noroozi, M.: Unsupervised video representation learning by bidirectional feature prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1670–1679 (2021)
5. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9922–9931 (2020)
6. Bucher, M., Herbin, S., Jurie, F.: Semantic bottleneck for computer vision tasks. In: Asian Conference on Computer Vision. pp. 695–712. Springer (2018)
7. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision. pp. 132–149 (2018)
8. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
10. Chen, B., Selvaraju, R.R., Chang, S.F., Niebles, J.C., Naik, N.: Previts: Contrastive pretraining with video tracking supervision. arXiv preprint arXiv:2112.00804 (2021)
11. Chen, P., Huang, D., He, D., Long, X., Zeng, R., Wen, S., Tan, M., Gan, C.: Rspnet: Relative speed perception for unsupervised video representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 1 (2021)
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
13. Chen, Z., Bei, Y., Rudin, C.: Concept whitening for interpretable image recognition. Nature Machine Intelligence **2**(12), 772–782 (2020)
14. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: NIPS. vol. 2, p. 4 (2013)
15. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: Tclr: Temporal contrastive learning for video representation. arXiv preprint arXiv:2101.07974 (2021)
16. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)
17. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H.: Motion-aware contrastive video representation learning via foreground-background

merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9716–9726 (2022)
18. Ding, S., Qian, R., Xiong, H.: Dual contrastive learning for spatio-temporal representation. arXiv preprint arXiv:2207.05340 (2022)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
20. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Advances in neural information processing systems **27** (2014)
21. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: International Conference on Machine Learning. pp. 3015–3024. PMLR (2021)
22. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3299–3309 (2021)
23. Gao, P., Lu, J., Li, H., Mottaghi, R., Kembhavi, A.: Container: Context aggregation network. arXiv preprint arXiv:2106.01401 (2021)
24. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
25. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010)
26. Han, T., Xie, W., Zisserman, A.: Video representation learning by dense predictive coding. In: Proceedings of the IEEE international conference on computer vision Workshops. pp. 0–0 (2019)
27. Han, T., Xie, W., Zisserman, A.: Memory-augmented dense predictive coding for video representation learning. In: Proceedings of the European conference on computer vision. pp. 312–329. Springer (2020)
28. Han, T., Xie, W., Zisserman, A.: Self-supervised co-training for video representation learning. arXiv preprint arXiv:2010.09709 (2020)
29. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
30. Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., Dou, D.: Discriminative sounding objects localization via self-supervised audiovisual matching. Advances in Neural Information Processing Systems **33**, 10077–10087 (2020)
31. Huang, D., Wu, W., Hu, W., Liu, X., He, D., Wu, Z., Wu, X., Tan, M., Ding, E.: Ascnet: Self-supervised video representation learning with appearance-speed consistency. arXiv preprint arXiv:2106.02342 (2021)
32. Huang, L., Liu, Y., Wang, B., Pan, P., Xu, Y., Jin, R.: Self-supervised video representation learning by context and motion decoupling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13886–13895 (2021)
33. Jabri, A., Owens, A., Efros, A.A.: Space-time correspondence as a contrastive random walk. arXiv preprint arXiv:2006.14613 (2020)

34. Jenni, S., Jin, H.: Time-equivariant contrastive video representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9970–9980 (2021)
35. Jenni, S., Meishvili, G., Favaro, P.: Video representation learning by recognizing temporal transformations. In: Proceedings of the European conference on computer vision. pp. 425–442. Springer (2020)
36. Kim, D., Cho, D., Kweon, I.S.: Self-supervised video representation learning with space-time cubic puzzles. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8545–8552 (2019)
37. Kim, D., Cho, D., Yoo, D., Kweon, I.S.: Learning image representations by completing damaged jigsaw puzzles. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 793–802. IEEE (2018)
38. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International Conference on Machine Learning. pp. 5338–5348. PMLR (2020)
39. Kuang, H., Zhu, Y., Zhang, Z., Li, X., Tighe, J., Schwertfeger, S., Stachniss, C., Li, M.: Video contrastive learning with global context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3195–3204 (2021)
40. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
41. Li, R., Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T.: Motion-focused contrastive learning of video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2105–2114 (2021)
42. Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., Yang, M.H.: Joint-task self-supervised learning for temporal correspondence. arXiv preprint arXiv:1909.11895 (2019)
43. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of the European conference on computer vision. pp. 513–528 (2018)
44. Liu, X., Qian, R., Zhou, H., Hu, D., Lin, W., Liu, Z., Zhou, B., Zhou, X.: Visual sound localization in the wild by cross-modal interference erasing. arXiv preprint arXiv:2202.06406 **2** (2022)
45. Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., Zhou, X., Wu, W., Dai, B., Zhou, B.: Learning hierarchical cross-modal association for co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10462–10472 (2022)
46. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. arXiv preprint arXiv:2201.07786 (2022)
47. Losch, M., Fritz, M., Schiele, B.: Interpretability beyond classification output: Semantic bottleneck networks. arXiv preprint arXiv:1907.10882 (2019)
48. Luo, D., Liu, C., Zhou, Y., Yang, D., Ma, C., Ye, Q., Wang, W.: Video cloze procedure for self-supervised spatio-temporal learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11701–11708 (2020)
49. Luo, Z., Peng, B., Huang, D.A., Alahi, A., Fei-Fei, L.: Unsupervised learning of long-term motion dynamics for videos. In: Proceedings of the IEEE international conference on computer vision. pp. 2203–2212 (2017)
50. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceed-

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020)
51. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. pp. 527–544. Springer (2016)
52. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
53. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11205–11214 (2021)
54. Piergiovanni, A., Angelova, A., Ryoo, M.S.: Evolving losses for unsupervised video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 133–142 (2020)
55. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: European Conference on Computer Vision. pp. 292–308. Springer (2020)
56. Qian, R., Li, Y., Yuan, L., Gong, B., Liu, T., Brown, M., Belongie, S., Yang, M.H., Adam, H., Cui, Y.: Exploring temporal granularity in self-supervised video representation learning. arXiv preprint arXiv:2112.04480 (2021)
57. Qian, R., Li, Y., Liu, H., See, J., Ding, S., Liu, X., Li, D., Lin, W.: Enhancing self-supervised video representation learning via multi-level feature optimization. arXiv preprint arXiv:2108.02183 (2021)
58. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. arXiv preprint arXiv:2008.03800 (2020)
59. Recasens, A., Luc, P., Alayrac, J.B., Wang, L., Strub, F., Tallec, C., Malinowski, M., Pătrăucean, V., Altché, F., Valko, M., et al.: Broaden your views for self-supervised video learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1255–1265 (2021)
60. Regatti, J.R., Deshmukh, A.A., Manavoglu, E., Dogan, U.: Consensus clustering with unsupervised representation learning. arXiv preprint arXiv:2010.01245 (2020)
61. Sawada, Y., Nakamura, K.: Concept bottleneck model with additional unsupervised concepts. arXiv preprint arXiv:2202.01459 (2022)
62. Seel, N.M.: Encyclopedia of the Sciences of Learning. Springer Science & Business Media (2011)
63. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
64. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743 (2019)
65. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 776–794. Springer (2020)
66. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
67. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

68. Wang, J., Jiao, J., Bao, L., He, S., Liu, W., Liu, Y.h.: Self-supervised video representation learning by uncovering spatio-temporal statistics. arXiv preprint arXiv:2008.13426 (2020)
69. Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W.: 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. pp. 4006–4015 (2019)
70. Wang, J., Jiao, J., Liu, Y.H.: Self-supervised video representation learning by pace prediction. In: Proceedings of the European conference on computer vision (2020)
71. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
72. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)
73. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3024–3033 (2021)
74. Weinzaepfel, P., Lucas, T., Larlus, D., Kalantidis, Y.: Learning super-features for image retrieval. arXiv preprint arXiv:2201.13182 (2022)
75. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
76. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision. pp. 305–321 (2018)
77. Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., Hu, H.: Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16684–16693 (2021)
78. Xiong, S., Tan, Y., Wang, G.: Explore visual concept formation for image classification. In: International Conference on Machine Learning. pp. 11470–11479. PMLR (2021)
79. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10334–10343 (2019)
80. Yang, C., Xu, Y., Dai, B., Zhou, B.: Video representation learning with visual tempo consistency. arXiv preprint arXiv:2006.15489 (2020)
81. Yao, T., Zhang, Y., Qiu, Z., Pan, Y., Mei, T.: Seco: Exploring sequence supervision for unsupervised representation learning. arXiv preprint arXiv:2008.00975 (2020)
82. Yao, Y., Liu, C., Luo, D., Zhou, Y., Ye, Q.: Video playback rate perception for self-supervised spatio-temporal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6548–6557 (2020)
83. Yuan, L., Qian, R., Cui, Y., Gong, B., Schroff, F., Yang, M.H., Adam, H., Liu, T.: Contextualized spatio-temporal contrastive learning with self-supervision. arXiv preprint arXiv:2112.05181 (2021)