

Dual Contrastive Learning for Spatio-temporal Representation

Shuangrui Ding
dsr1212@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Rui Qian
qr021@ie.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong, China

Hongkai Xiong*
xionghongkai@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

ABSTRACT

Contrastive learning has shown promising potential in self-supervised spatio-temporal representation learning. Most works naively sample different clips to construct positive and negative pairs. However, we observe that this formulation inclines the model towards the background scene bias. The underlying reasons are twofold. First, the scene difference is usually more noticeable and easier to discriminate than the motion difference. Second, the clips sampled from the same video often share similar backgrounds but have distinct motions. Simply regarding them as positive pairs will draw the model to the static background rather than the motion pattern. To tackle this challenge, this paper presents a novel dual contrastive formulation. Concretely, we decouple the input RGB video sequence into two complementary modes, static scene and dynamic motion. Then, the original RGB features are pulled closer to the static features and the aligned dynamic features, respectively. In this way, the static scene and the dynamic motion are simultaneously encoded into the compact RGB representation. We further conduct the feature space decoupling via activation maps to distill static- and dynamic-related features. We term our method as **Dual Contrastive Learning for spatio-temporal Representation (DCLR)**. Extensive experiments demonstrate that DCLR learns effective spatio-temporal representations and obtains state-of-the-art or comparable performance on UCF-101, HMDB-51, and Diving-48 datasets.

CCS CONCEPTS

• Computing methodologies → Activity recognition and understanding; • Information systems → Video search.

KEYWORDS

Self-supervised Learning; Action Recognition

ACM Reference Format:

Shuangrui Ding, Rui Qian, and Hongkai Xiong. 2022. Dual Contrastive Learning for Spatio-temporal Representation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3503161.3547783>

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '22, October 10–14, 2022, Lisboa, Portugal

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3547783>

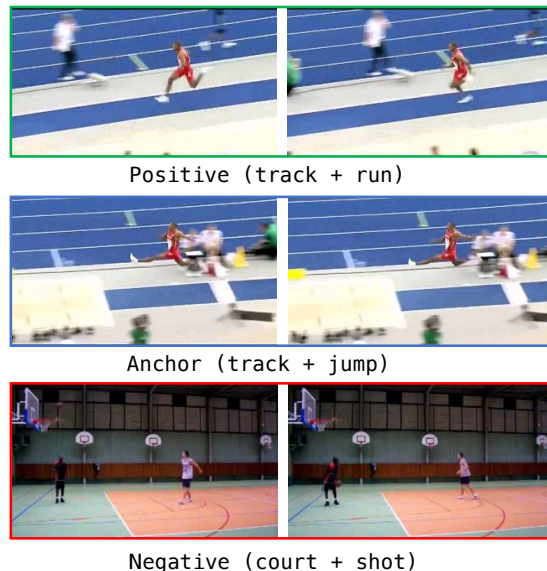


Figure 1: An illustration for positive and negative pair in spatio-temporal contrastive learning. For positive pairs, the two clips have the same background (track) but distinct motions (run vs jump). And for the negative pair, the difference in the background scene (track vs court) is much more noticeable than the difference in motion (jump vs shot). These two phenomena cause static scene bias.

1 INTRODUCTION

Recently, self-supervised spatio-temporal representation learning has attracted great interest in the computer vision community. Compared with traditional supervised settings, the core of this learning scheme is to extract general representations from large-scale video data without resorting to human annotations. Instead of supervision from the costly manual labels, self-supervised learning obtains supervision from the unlabeled data themselves, enabling the utilization of millions of freely accessible videos on the Internet.

Inspired by the success in image domain [7, 18], contrastive-based methods have been expanded to spatio-temporal representation learning [12, 42] and achieved superior performance compared to previous pretext task-based methods [26, 38, 59]. In particular, a common implementation of spatio-temporal contrastive learning is to sample two temporally different clips from each video, then regard pairs from the same (different) video as positive (negative) samples. The model is forced to draw ‘positive’ pairs closer in the feature space and push apart the ‘negative’ pairs. Under this formulation, the final representations are encouraged to capture the

discriminative information, which greatly facilitates the performance. However, previous works [9, 51] reveal that this learning diagram tends to favor the static cues while focusing less on the motion. The potential reason lies in two aspects. First, the static scene bias exists in the positive pair formulation. In vanilla contrastive learning, we sample different temporal clips from each video. Those clips usually share a similar background but have a subtle difference in motion. As shown in Fig. 1, the positive pair owns the same background (track field) but different motions (jump vs run). If the model is encouraged to pull such biased positive pairs closer in the feature space, the model will naturally attend to the characteristics of the background scene but fail to capture the motion information. And for negative pairs, the background information still appears more salient than the motions. Take the negative pair in Fig. 1 for instance, the scenes of the indoor basketball court and the outdoor track field almost dominate the entire screen. Thus, the considerable distinction between these two scenes seems to be sufficient for the model to push the negative pair away while the difference in motion patterns in the negative pair is hard to get noticed. To verify whether the background shortcut truly exists in contrastive learning, we utilize the static frame, which does not carry motion information, for analysis. We regard the static frame as the positive sample of the original RGB clips to train the model. We find the model pretrained in this manner can achieve almost the same performance as the vanilla method (48.1% vs 48.9%). This empirical observation demonstrates that pulling RGB video pairs closer is basically equivalent to pulling the RGB video and the static frame closer, which reveals vanilla contrastive learning degrades the model to focus on background cues and learn static-biased representations. More pieces of evidence are shown in Table 4.

Hence, here comes a question on how to formulate the learning scheme that makes contrastive learning take both scene and motion into consideration. To delve into this problem, our paper proposes a novel approach named **Dual Contrastive Learning for spatio-temporal Representation (DCLR)**. We perceive scene and motion as two kinds of orthogonal and complementary information sources and decouple them at two levels. We first decouple the static and dynamic information on the inputs and define a dual contrastive-based objective that enables the model to capture both static and dynamic features in video data. Particularly, given an RGB video sequence, we repeat a random frame along the temporal axis as the static data and regard the frame difference as the dynamic data. We train the model by respectively minimizing two dual contrastive losses. One is the alignment between the original RGB sequence and static input, the other is between RGB and dynamic input. In addition, to emphasize motion learning, we mine the truly aligned motion positive pair of the RGB sequence across different videos. We do not view the clips from the same video as corresponding motion positive pairs. Rather, we maintain a dynamically updated feature extractor and retrieve the most similar motions as the corrected positive pairs. Besides the static-dynamic decoupling in the input space, we further enhance the decoupling in the feature space. Specifically, we constrain that the RGB sequence feature activations should be consistent with the combination of static frame and frame difference activations. Meanwhile, we use the latter two activation maps to refine the static and dynamic related features in the RGB

representation for dual contrastive learning. We evaluate the proposed DCLR on three action recognition downstream benchmarks, UCF-101, HMDB-51, and Diving-48, and manifest the effectiveness of each component in our framework. The experimental results demonstrate that DCLR enables contrastive spatio-temporal representation learning to resist the background shortcuts and achieve better generalization ability.

To sum up, our contributions are as follows:

- We formulate a novel self-supervised learning scheme, dual contrastive learning, motivated by the static bias in the spatio-temporal representation learning.
- We decouple static and dynamic cues in both data input and feature space to enhance dual contrastive learning.
- We achieve state-of-the-art or competitive results on downstream action recognition and video retrieval across UCF-101, HMDB-51, and Diving-48 datasets.

2 RELATED WORK

Self-supervised Representation Learning. The target of representation learning is to learn a transferable encoder that extracts desired characteristics from input data and filters redundant information. In traditional supervised learning like classification, [19, 47, 58] directly use the category labels as the learning objective. While in self-supervised learning, it is nontrivial to define the objective. Early works design various pretext tasks, e.g., rotation [13], colorization [27], jigsaw [10, 37], to learn certain attributes. But the handcrafted pretext tasks are limited in performance. Later, contrastive learning promotes great progress in image representation learning [39, 60, 61]. It employs the consistency between multiple views of the same instance as self-supervisory signal [7, 46]. Compared with the supervised settings, the multi-view constraint provides richer information, including semantics and some other instance-specific characters [62, 67]. Therefore, contrastive learning can obtain comprehensive representations that preserve unique information of each instance. However, in the video domain, due to the complex spatio-temporal structures, the naive multi-view constraint is far from satisfactory. In this work, we formulate a decoupled multi-view constraint to better fit the spatio-temporal nature.

Spatio-temporal Representation Learning. Inspired by self-supervised learning in image domain, a line of works relies on pretext tasks to learn spatio-temporal representations. Since videos contain much richer characteristics, there are more pretext tasks: temporal ordering [38, 59, 63], spatio-temporal puzzles [26, 52], playback speed prediction [4, 24], temporal cycle-consistency [23, 33, 54], and future prediction [3, 15, 16, 36, 48, 49]. Besides, some works expand the contrastive learning pipeline to the video domain by sampling different clips or modalities to formulate multi-view constraint [1, 12, 17, 42, 43, 53]. Recent works [9, 51] observe that vanilla contrastive-based methods lead to static scene bias and attend less to temporal dynamic. To deal with it, our work reformulates the conventional contrastive learning as a dual learning problem, which encodes scene-debiased and motion-aware representations. [22, 66] also adopt dual form of contrastive learning in video data. Specifically, [66] samples clips of the same timestamp for spatial contrast and samples clips of different timestamps

for temporal contrast to separately learn spatial and temporal attributes. Since there is motion misalignment in temporally different clips, the static bias could still exist in [66]. While ours can avoid this problem by directly maximizing the agreement between video clips with corresponding temporally aligned positive pairs. And for [22], they also conduct contrastive learning via three streams to extract RGB, static and dynamic features. But [22] only decouples representations at the level of data input. We extend this decoupling methodology to feature space. We adopt the attention map to refine the static-related and dynamic-related contrasts and restrict the complementarity of RGB features. The superior results compared to [22, 66] manifest the effectiveness of our dual formulations.

Elimination of Background Bias in Video. The exploration of mitigating the background bias [8, 20, 34, 55] in the video emerges for a long time. [8] proposes to mitigate scene bias by augmenting the standard cross-entropy loss, where it leverages human-masked-out videos to tell model whether the video contains action. [34] aims at alleviating static representation bias in existing datasets, where proposes a procedure to reassemble existing datasets. In self-supervised domain, other modality like optical flow [32, 57] are employed to emphasize motion information explicitly. To utilize the motion information implicitly in RGB, DSM [50] and BE [51] decouple the motion and context by deliberately constructing the motion-aware positive/negative samples through disturbance. Moreover, FAME [9] proposes a copy-paste augmentation technique to keep the foreground motion intact. It copies the foreground area onto the other background to resist the background shortcut. In this paper, we solve this bias problem from two perspectives. On the one hand, we decouple RGB into the easily accessible static frame and frame difference and train the model via dual contrastive loss. On the other hand, we sample corresponding motion patterns across videos in a large pool to weaken the bias in positive pair sampling.

3 APPROACH

In this section, we elaborate on our framework of dual contrastive learning. Our goal is to learn compact and rich feature representations from videos, which are not only scene-related but also contain temporal dynamic information. We first revisit vanilla spatio-temporal contrastive learning in Sec. 3.1. And in Sec. 3.2, we illustrate the construction of positive (negative) pairs in dual contrastive learning. Sec. 3.3 introduces the further decoupling in future space via activation maps. Finally, we give the full objective in Sec. 3.4.

3.1 Spatio-temporal contrastive learning

The vanilla spatio-temporal contrastive learning adopts instance discrimination [56] in a fully self-supervised manner. For ease of notation, we represent video data $v \in \mathbb{R}^{C \times T \times H \times W}$, where C, T, H, W denote the dimensions of the channel, timespan, height, width, respectively. We notate the video encoder as $f: \mathbb{R}^{C \times T \times H \times W} \rightarrow \mathbb{R}^D$, where D is the dimension of representation. Similar to image domain [7, 14, 18], it maximizes the similarity between two different views v_i, v_j of one query sample v and minimizes the similarity between negative pairs. Two different views of one video are sampled from two timestamps and then are processed by the temporal-consistent augmentation to reserve motion information [12, 42]. We simply adopt other samples from the mini-batch as negative

sample. Assuming the number of mini-batch is N , we take rest $2(N-1)$ examples in same mini-batch as negative samples. Having the positive pairs (v_i, v_j) , the loss function $\mathcal{L}_{VV}$ is formulated as

$$\mathcal{L}_{VV} = I(f(v_i); f(v_j)) + I(f(v_j); f(v_i)), \quad (1)$$

$$I(f(v_i); f(v_j)) = -\log \frac{\exp(\text{sim}(f(v_i), f(v_j))/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\text{sim}(f(v_i), f(v_k))/\tau)}, \quad (2)$$

where $\mathbb{I}_{[k \neq i]} \in \{0, 1\}$ is an indicator function and τ is the temperature hyper-parameter. $\text{sim}(z_i, z_j)$ measures the cosine similarity between the latent representation, i.e., $\text{sim}(z_i, z_j) = z_i^T z_j / (\|z_i\|_2 \|z_j\|_2)$.

As mentioned in the Sec. 1, this contrastive formulation possess severe background bias [9, 51]. Two temporally different clips usually own similar static backgrounds but slightly diverse motions. In this way, contrastive learning would intuitively prioritize the background information alignment rather than motion. Such bias contributes to the weak generalization ability of the model.

3.2 Static-dynamic decoupling in data input

To mitigate the aforementioned background bias, we decouple the original input to perform dual contrastive learning, where we split the original video data into static frame and frame difference.

Static Frame. We define static frame $s \in \mathbb{R}^{C \times T \times H \times W}$. Simply, we repeat a randomly selected frame to carry the static information without any dynamic motion. Mathematically,

$$s = \underbrace{[v_t, \dots, v_t]}_{T \text{ times}} \quad \text{for any } t \in [1, T] \quad (3)$$

where $v \in \mathbb{R}^{C \times T \times H \times W}$ is the original clip and t is the index of temporal dimension.

Frame Difference. We denote frame difference as $d \in \mathbb{R}^{C \times T \times H \times W}$. By differentiating adjacent frames iteratively, frame difference conveys natural motion information where moving areas possess a great magnitude and static cues are eliminated in this data input. Mathematically,

$$d = v_{2:T+1} - v_{1:T}, \quad (4)$$

where $v \in \mathbb{R}^{C \times (T+1) \times H \times W}$ is the original clip and subscript is the index of temporal dimension. Except for the frame difference, we also consider optical flow to convey motion information. Though the quality of optical flow looks more accurate, the extraction of the dense optical flow produces an unaffordable computational cost. Therefore, we adopt frame difference as a cheaper substitute.

Given complementary modalities s and d , we can losslessly recover v with simple union operation $s \cup d$. To this end, it is feasible to encode the feature of v containing both characteristics of s and d at the same time. Therefore, we transform the vanilla contrastive objective $\mathcal{L}_{VV}$ into a dual form as

$$\mathcal{L}_{VV} \rightarrow \mathcal{L}_{VS} + \mathcal{L}_{VD}, \quad (5)$$

$$\mathcal{L}_{VS} = I(f(v_i); f(s_j)) + I(f(v_j); f(s_i)), \quad (6)$$

$$\mathcal{L}_{VD} = I(f(v_i); f(d_j)) + I(f(v_j); f(d_i)), \quad (7)$$

where the function $I(\cdot; \cdot)$ is the same as Eq. 2. $\mathcal{L}_{VS}$ ($\mathcal{L}_{VD}$) optimizes the alignment between video $f(v)$ and $f(s)$ ($f(d)$). By minimizing this dual loss term, $f(v)$ should contain both static and dynamic characters.

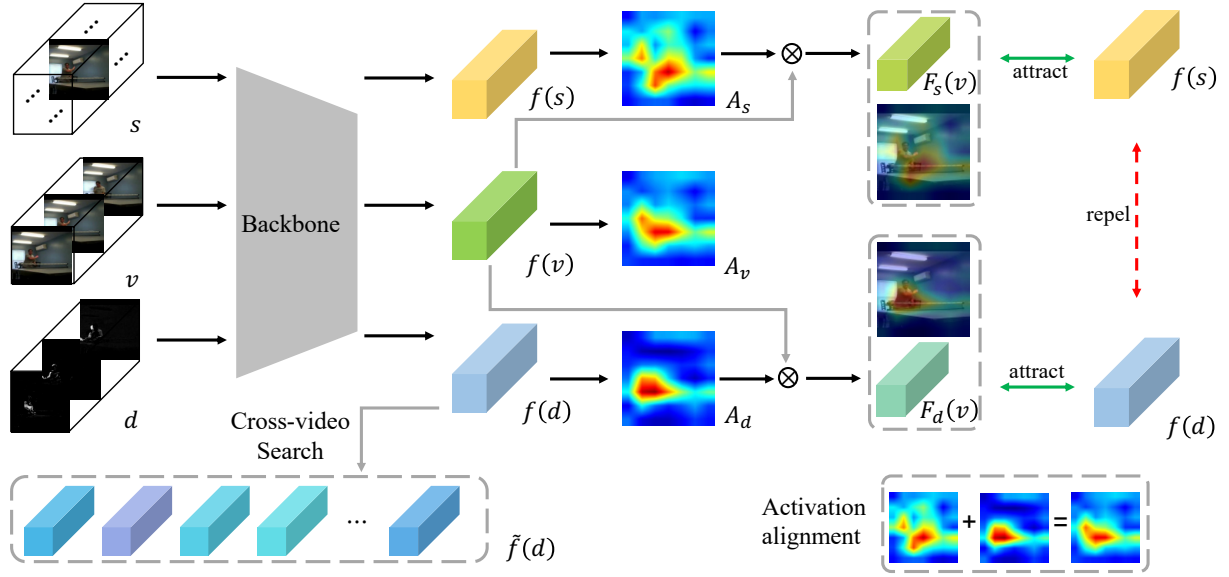


Figure 2: An overview of the proposed method. We first feed three data inputs s, v, d into the backbone. We search the corresponding motion patterns in the large sample pool established by $\tilde{f}(d)$ to correct the motion misalignment in the positive sample formulation. We utilize the activation maps as a concrete referer to purify static and dynamic features, and employ the consistency constraint to let $f(v)$ cover the joint of $f(s)$ and $f(d)$. *Best viewed in color.*

However, one concern in the decoupled contrastive optimization process is that $f(v)$ possibly collapses to the intersection of $f(s)$ and $f(d)$. In other words, $f(s)$ and $f(d)$ do not contain desired complementary information but only carry limited common information. This collapse exactly goes against our goal. Hence, in order to prevent the potential collapse problem, we enforce two decoupled representations orthogonal by maximizing a regularization term $\mathcal{L}_{SD}$, i.e.,

$$\mathcal{L}_{SD} = I(f(s_i); f(d_i)) + I(f(s_j); f(d_j)). \quad (8)$$

Note that in contrast to $\mathcal{L}_{VS}$ and $\mathcal{L}_{VD}$ which take two different views of one video data, we use static frame and frame difference from the same view to ensure disparity between $f(s)$ and $f(d)$. The intuition is that if the model learns to attract positive pairs from the same view, it will easily attend to redundant information, which deviates from the vital characteristics. By maximizing $\mathcal{L}_{SD}$, we could guarantee the complementarity between $f(s)$ and $f(d)$, and let $f(v)$ simultaneously include much static as well as dynamic characteristics that help unbiased video understanding. We provide empirical results to support this formulation in the ablation study.

Now we have addressed the collapse problem by introducing the regularization term $\mathcal{L}_{SD}$, but the static bias in dual contrastive formulation remains. Concretely, the positive pair often shares similar backgrounds but differs in motions, i.e., given two views v_i and v_j , the corresponding dynamic motions d_i and d_j do not always align. Hence, if we directly pull $f(v_i)$ and $f(d_j)$ closer through $\mathcal{L}_{VD}$, the model cannot learn helpful dynamic-related knowledge. Therefore, we conduct the cross-video search to figure out the truly aligned motion patterns as corrected positive motion pairs. Motivated by the queue mechanism in MoCo [18], we maintain a slowly changing frame difference feature extractor $\tilde{f}$ which is

updated every few epochs, and a queue of length L to store extracted frame difference features. In each iteration, given two-view frame difference input d_i and d_j , we update the queue with $\tilde{f}(d_j)$, and apply $\tilde{f}(d_i)$ as query to retrieve similar pairs in the memory queue.

For a neat presentation, we omit the subscript and denote the query as (v, d) . We calculate the cosine similarity $S \in \mathbb{R}^L$ between $f(d)$ and each frame difference feature in the queue. In this way, S reveals the pair-wise similarity in dynamic motions. Hence, we can intuitively obtain cross-video motion pattern correspondence. We consider several variants of implementations in the retrieval stage. One simplest way is to retrieve the sample with highest similarity score, denoted as $(\tilde{v}, \tilde{d})$. Then we use $(v, \tilde{d})$ and $(\tilde{v}, d)$ to form modified $\mathcal{L}_{VD}$, i.e.,

$$\mathcal{L}_{VD} = I(f(v); f(\tilde{d})) + I(f(\tilde{v}); f(d)). \quad (9)$$

Besides, we can select a subset of samples with topK highest similarity score $[(\tilde{v}_1, \tilde{d}_1), (\tilde{v}_2, \tilde{d}_2), \dots, (\tilde{v}_K, \tilde{d}_K)]$, where K is the number of samples in the selected subset. Meanwhile, we employ the cosine similarity score as prior knowledge. Then the loss calculation equals to

$$\mathcal{L}_{VD} = \sum_{i=1}^K p_i [I(f(v); f(\tilde{d}_i)) + I(f(\tilde{v}_i); f(d))] \quad (10)$$

where $p_i = S[i] / \sum_{i=1}^K S[i]$ and we sort S in descending order. Cross-instance retrieval is a common and effective way to align the high-level semantics shown in recent works [11, 29]. In our work, the motivation for retrieving cross-video samples is different, where we aim to mitigate the motion pattern misalignment in positive pairs. Through this process, we fully leverage the natural characters of frame difference as well as the learned knowledge to mitigate

the bias in the dual contrastive formulation. And we also explore reformulating $\mathcal{L}_{VS}$ in a similar manner. There is no significant gain in performance, which is concordant with our motivation that there only exists dynamic misalignment in positive pairs. The detailed discussions are displayed in the ablation study.

3.3 Static-dynamic decoupling in feature space

We further enhance static-dynamic decoupling in feature space. To do this, we consider an abstract measurement of the high-level features, which is the activation maps [2, 68]. The activation maps involve richer information on the spatio-temporal distribution of the extracted features. Particularly, we obtain the class-agnostic activation maps [2] by calculating the summation of the feature maps $F(\cdot) \in \mathbb{R}^{C \times T \times H \times W}$ over the channel dimension. For example, the activation $A_v \in \mathbb{R}^{T \times H \times W}$ for $f(v)$ is

$$A_v[t, h, w] = \sum_{c=1}^C |F(v)|[c, t, h, w], \quad (11)$$

where t, h, w denotes spatio-temporal index, C is channel dimension. A_s and A_d are computed in the same fashion.

Naturally, A_d attends more to the spatio-temporal areas that contain dynamic motions, while A_s focuses on areas with discriminative static cues. Inspired by this, the activation A_v for $f(v)$ should jointly highlight scene and motion related areas. Thus, we derive the activation alignment constraint:

$$\mathcal{L}_{ac} = \|A_v - (A_s + A_d)\|_1. \quad (12)$$

We apply min-max normalization over the whole spatio-temporal dimensions to A_v and $(A_s + A_d)$ in loss calculation. To stabilize the training, we only backpropagate the gradient of $\mathcal{L}_{ac}$ to $f(v)$ stream and stop the gradient to $f(s)$ or $f(d)$ stream. In this way, A_v is encouraged to cover both dynamic and static reference areas and avoid falling into a trivial solution.

Besides the direct alignment in activation maps, we rely on A_s and A_d to purify the static and dynamic related features of $f(v)$. Particularly, we perform global weighted pooling on $F(v)$ to obtain the refined features. Given A_d , we represent dynamic related features $f_d(v)$ as

$$f_d(v) = \frac{\sum_{t,h,w} F(v)[t, h, w] \cdot A_d[t, h, w]}{\sum_{t,h,w} A_d[t, h, w]}. \quad (13)$$

Static related features $f_s(v)$ are obtained similarly. We now replace vanilla $f(v)$ with decoupled features in the dual formulation, i.e.,

$$I(f(v); f(s)) \rightarrow I(f_s(v); f(s)), \quad (14)$$

$$I(f(v); f(d)) \rightarrow I(f_d(v); f(d)). \quad (15)$$

Through feature space decoupling, we filter out the possible noise that may interfere with our dual contrastive objective and enhance the representation ability.

3.4 The Full Objective

The ultimate framework is illustrated in Fig. 2. We first feed three data types s, v, d into the video encoder f . It is worth noting that we adopt the same backbone for RGB, static frame, and frame difference since we empirically find that using separate backbones for three inputs achieves similar results with the same backbone. The underlying reason might be after the pre-processing normalization, the

distribution of the three data types s, v, d is not that different. Thus we adopt the same backbone to reduce parameters and training costs. For the dynamic motion branch, we maintain a dynamically updated feature extractor $\tilde{f}$ to establish cross-video correspondence and correct the motion misalignment in the original positive pair formulation. Then, we use activation maps A_s and A_d , as soft masks to decouple static scene and dynamic motion related features for dual contrastive learning, and leverage the consistency between A_v and $A_s + A_d$ to enhance the agreement between $f(v)$ and the union $f(s) \cup f(d)$.

Overall, the training loss consists of two parts, the dual contrastive term and the activation alignment term:

$$\mathcal{L} = (\mathcal{L}_{VS} + \mathcal{L}_{VD} - \mathcal{L}_{SD}) + \lambda \mathcal{L}_{ac}, \quad (16)$$

where λ is the balancing hyper-parameter, set to 0.5 in default. Considering that the cross-video retrieval as well as the inferred A_s and A_d are not reliable in early training stage, we perform truly aligned motion positive pair correction in Eq. 9 and feature refinement in Eq. 14 & 15 after a few epochs.

4 EXPERIMENTS

4.1 Implementation Details

Dataset. We use four popular video benchmarks for experiments, Kinetics-400 [5], UCF-101 [44], HMDB-51 [31] and Diving-48 [34]. **Kinetics-400** [5] is a large-scale and high-quality dataset for action recognition, collected from realistic YouTube videos. Kinetics-400 contains over 240K video clips of 400 action classes. **UCF-101** [44] is an action recognition dataset consisting of over 13k clips covering 101 action classes. **HMDB-51** [31] is another action recognition dataset with 51 action categories and around 7,000 annotated clips. **Diving-48** [34] involves 18k diving clips of 48 fine-grained diving categories, which majorly vary in motions and are no vast difference in the scenes. We pretrain the model on the training set of UCF-101 or Kinetics-400, and evaluate on the split 1 of UCF-101 and HMDB-51, or the V2 test set of Diving-48.

Self-supervised Pretraining. We use R(2+1)D-18 [47], with 14.4M parameters, as the video encoder, and share the same network to extract RGB, static frame, and frame difference features. We randomly sample two temporally different clips in each video as two views and apply temporally consistent random resized crop and random horizontal flip to obtain the frame sequence. We decouple the static frame input s and the frame difference input d as described in Eq. (3) & (4). We apply color jitter and Gaussian blur to augment the RGB input v , static frame input s . In this way, v, s and d are all of spatio-temporal resolution $16 \times 112 \times 112$ and input to the same encoder. We pretrain the model for 200 epochs on UCF-101 or 100 epochs on Kinetics-400. An SGD optimizer is adopted with the initial learning rate of 10^{-2} and weight decay of 10^{-4} . For the cross-video search, on UCF-101, we update the slowly changed feature extractor every 10 epochs with a queue length of 2048. On Kinetics-400, we update the extractor every 5 epochs with a queue length of 16384.

Action Recognition. We use the pretrained parameters except for the last fully-connected layer for initialization. We employ two popular protocols to validate the self-supervised representations: (1) *Finetune* the whole network with action labels; (2) Freeze the

Method	Backbone	Pretrain Dataset	Frames	Res.	Freeze	UCF-101	HMDB-51
CCL [28]	R3D-18	Kinetics-400	16	112	✓	52.1	27.8
MemDPC [16]	R3D-34	Kinetics-400	40	224	✓	54.1	30.5
RSPNet [6]	R3D	Kinetics-400	16	112	✓	61.8	42.8
MLRep [41]	R3D	Kinetics-400	16	112	✓	63.2	33.4
FAME [9]	R(2+1)D	Kinetics-400	16	112	✓	72.2	42.2
DCLR(Ours)	R(2+1)D	Kinetics-400	16	112	✓	72.3	46.4
VCP [35]	R3D	UCF-101	16	112	✗	66.3	32.2
IIC [45]	C3D	UCF-101	16	112	✗	72.7	36.8
MLRep [41]	R3D	UCF-101	16	112	✗	76.2	41.1
TempTrans [24]	R(2+1)D	UCF-101	16	112	✗	81.6	46.4
DCLR(Ours)	R(2+1)D	UCF-101	16	112	✗	82.3	50.1
3DRotNet [25]	R3D	Kinetics-400	16	112	✗	62.9	33.7
Pace Prediction [53]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
MemDPC [16]	R3D	Kinetics-400	40	224	✗	78.1	41.2
Pace [53]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
VideoMoCo [40]	R(2+1)D	Kinetics-400	32	112	✗	78.7	49.2
MLRep [41]	R3D	Kinetics-400	16	112	✗	79.1	47.6
TempTrans [24]	R3D	Kinetics-400	16	112	✗	79.3	49.8
RSPNet [6]	R(2+1)D	Kinetics-400	16	112	✗	81.1	44.6
ASCNet [21]	R3D	Kinetics-400	16	112	✗	80.5	52.3
SRTC [65]	R(2+1)D	Kinetics-400	16	112	✗	82.0	51.2
DCLR(Ours)	R(2+1)D	Kinetics-400	16	112	✗	83.3	52.7

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*.

Method	Pretrain Dataset	Res.	Top-1
Random Init.	-	-	50.7
BE [51]	UCF-101	224	58.8
FAME [9]	UCF-101	224	67.8
DCLR(Ours)	UCF-101	112	72.7
BE [51]	Kinetics-400	224	62.4
FAME [9]	Kinetics-400	224	72.9
DCLR(Ours)	Kinetics-400	112	75.1

Table 2: Top-1 accuracy on Diving-48 dataset. We compare different pretrain settings and evaluate on V2 labels.

backbone and train the last linear classifier, denoted as *linear probe*. During the inference phase, we follow the prevalent evaluation protocols [53, 59] to uniformly sample ten 16-frame clips from each video, then center crop and resize them to 112×112 . We average the prediction of each clip as video-level action prediction and report Top-1 accuracy to measure the action recognition performance.

Video Retrieval. We extract spatio-temporal features from the pretrained model without any further training. Based on the cosine similarity, videos in the test set retrieve the Top- k nearest neighbors in the training set. If the category of the test set exists in the k nearest neighbors, it counts as a hit. Following [53, 59], we average representations of ten uniformly sampled clips. We report Top- k recall $R@k$ for evaluation.

4.2 Evaluation on Downstream Tasks

Action Recognition. We first present action recognition on UCF-101 and HMDB-51 in Table 1. We report *linear probe* and *finetune* Top-1 accuracy. For a fair comparison, we do not include the works with different evaluation settings and much deeper backbone or non-single modality e.g., optical flow, audio, and text.

In *linear probe* settings, our method obtains the best result on both UCF-101 and HMDB-51. Even though MLRep [41] carefully devises the multi-level feature optimization and temporal modeling, DCLR beats MLRep [41] significantly, i.e., 9.1% and 13% improvement on UCF-101 and HMDB-51 respectively. In addition, DCLR surpasses FAME [9] by 4.2% on HMDB51, an approach that emphasizes motion pattern learning, demonstrating the effectiveness of our dual contrastive learning framework.

In *finetune* protocol, DCLR still achieves the best results among RGB-only methods. Note that [6, 21, 24, 65] introduce diverse temporal transformations or carefully design temporal pretext tasks. The superiority over them proves our method distills effective spatio-temporal representations by decoupling both levels of data input and feature space.

Additionally, in Table 2, we report finetune results on a more challenging Diving-48 dataset [34], where all videos share a similar background and only differ in long-term motion patterns. In Diving-48, the fine-grained categories are not strongly correlated with static backgrounds anymore. Thus, the result on such a motion-heavy dataset can better display whether the model captures the motion-aware representations. We compare our method with FAME [9], which applies motion inductive augmentation to highlight motion patterns in contrastive learning. Our method DCLR greatly outperforms FAME with $2\times$ smaller resolution on both UCF-101 and Kinetics-400 pretrain datasets. It demonstrates that DCLR considerably enhances long-range motion pattern modeling and solves the background bias in naive spatio-temporal contrastive learning.

Video Retrieval. We show the performance on video retrieval with $R@k$ in Table 3. All models are pretrained on UCF-101 for a fair comparison. We gain significant improvement on both UCF-101

Method	Backbone	UCF-101				HMDB-51			
		R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
VCP [35]	R3D	18.6	33.6	42.5	53.3	7.6	24.4	36.3	53.6
Pace [53]	R(2+1)D	25.6	42.7	51.3	61.3	12.9	31.6	43.2	58.0
PRP [64]	R(2+1)D	20.3	34.0	41.9	51.7	8.2	25.3	36.2	51.0
STS [52]	R3D	38.3	59.9	68.9	77.2	18.0	37.2	50.7	64.8
VCLR [30]	R2D-50	46.8	61.8	70.4	79.0	17.6	38.6	51.1	67.6
DCLR(Ours)	R(2+1)D	54.8	68.3	75.9	82.8	24.1	44.5	53.7	64.5

Table 3: Results on video retrieval task pretrained on UCF-101. We report R@k (k=1,5,10,20) on UCF-101 and HMDB-51.

$\mathcal{L}_{VV}$	$\mathcal{L}_{VS}$	$\mathcal{L}_{VD}$	$\mathcal{L}_{SD}$	Same-View	UCF-101
✓					48.9
	✓				48.1
		✓			55.2
	✓	✓			60.2
	✓	✓		✓	41.7
	✓	✓	✓		61.1
✓	✓				49.3
✓		✓			59.4
✓	✓	✓			61.7
✓	✓	✓	✓		62.4

Table 4: Top-1 linear probe accuracy on UCF-101. We compare different loss combinations, with $\mathcal{L}_{VV}$, $\mathcal{L}_{VS}$, $\mathcal{L}_{VD}$ to be minimized, $\mathcal{L}_{SD}$ to be maximized. The ‘same-view’ indicates whether adopting the same view samples for $\mathcal{L}_{VS}$ and $\mathcal{L}_{VD}$.

and HMDB-51 datasets. It indicates that DCLR formulation encodes static and dynamic characteristics into a more compact manifold.

4.3 Ablation Study

For further analysis, we dissect our approach on several crucial modules. We pretrain on UCF-101 for 200 epochs, and report the *linear probe* Top-1 accuracy.

Effectiveness of $\mathcal{L}_{VS}$, $\mathcal{L}_{VD}$, and $\mathcal{L}_{SD}$. To analyze the effectiveness of the decoupled learning objective, we make extensive preliminary experiments on various sets of loss. Note that in default settings, for $\mathcal{L}_{VV}$, $\mathcal{L}_{VS}$, and $\mathcal{L}_{VD}$, we use temporally different views of the same video to minimize the loss. Instead, for $\mathcal{L}_{SD}$ that needs to be maximized, we utilize the same view to avoid the collapse issue. We show the *linear probe* action recognition accuracy on UCF-101 in Table 4. Vanilla contrastive spatio-temporal representation learning baseline is located at the first row in the table, only using $\mathcal{L}_{VV}$. By comparing the baseline and various settings, we reach several important observations. **First**, considering the first three lines, adoption of $\mathcal{L}_{VS}$ gains a similar result with baseline (0.8%↓) while using $\mathcal{L}_{VD}$ significantly improves the performance (6.3%↑). It indicates that naively optimizing $f(v)$ via $\mathcal{L}_{VV}$ nearly equals pulling the video with its static frame, thus losing crucial dynamic motion characters. However, employing $\mathcal{L}_{VD}$ can resist the background shortcuts and boost the representation quality. The above observation strongly coincides with our motivation about the existing static bias in the contrastive formulation. **Second**, among the middle three lines, we find that jointly minimizing $\mathcal{L}_{VS}$ and $\mathcal{L}_{VD}$ improves the baseline by a large margin (11.3%↑), which means decoupling in data input can validly solve the background bias and help the model capture more compact and comprehensive

Setting	UCF-101	HMDB-51
None	64.3	37.7
Only $\tilde{f}(s)$	64.5	38.3
Only $\tilde{f}(d)$	67.1	40.1
Joint $\tilde{f}(s)$ & $\tilde{f}(d)$	67.2	39.8

Table 5: Ablation study on cross-video correspondence search. We compare the baseline and different search settings.

Num.	Dist.	UCF-101	HMDB-51
-	-	64.3	37.7
1	-	66.4	39.7
5	Uniform	66.9	39.5
5	Prior	67.1	40.1
10	Uniform	65.6	38.1
10	Prior	66.3	39.8

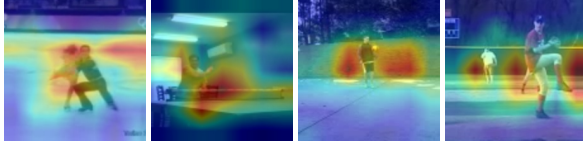
Table 6: Ablation study on cross-video retrieval sample distribution formulation. We compare the results with different numbers and distribution settings.

representations. And if we adopt the same-view setting to formulate $\mathcal{L}_{VS}$ and $\mathcal{L}_{VD}$, the performance will drop dramatically (7.2%↓). It is consistent with our hypothesis that the same view causes the model to deviate from the high-level semantic space and attend to redundant low-level information. Besides, the introduction of the regularization term $\mathcal{L}_{SD}$ further improves the accuracy by 0.9% compared to only utilization of $\mathcal{L}_{VS}$ and $\mathcal{L}_{VD}$. **Third**, we investigate the combination of $\mathcal{L}_{VV}$ and decoupled contrastive losses in the last four lines. Unsurprisingly, integrating $\mathcal{L}_{VV}$ and $\mathcal{L}_{VS}$ seems to bring no improvement (0.4%↑), while jointly optimizing $\mathcal{L}_{VV}$ and $\mathcal{L}_{VD}$ greatly enhances the performance (10.5%↑). Also, incorporating $\mathcal{L}_{VV}$ into the combination of $\mathcal{L}_{VS}$ and $\mathcal{L}_{VD}$ only leads to marginal improvement (60.2% vs 61.7%). The phenomenon is reproducible when we add $\mathcal{L}_{SD}$ at the same time (61.1% vs 62.4%). It is shown that $\mathcal{L}_{VS}$ and $\mathcal{L}_{VD}$ are sufficient for dual contrastive learning. Hence, $\mathcal{L}_{VV}$ is not imported in default setting.

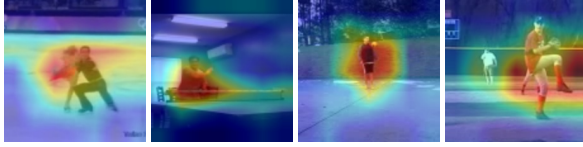
Cross-video Correspondence Search. We first compare different settings for cross-video correspondence search in Table 5. We observe that employing $\tilde{f}(d)$ to search similar motion patterns considerably facilitates the performance while using $\tilde{f}(s)$ only spurs marginal improvement. It proves that there exists a severe static bias in the original positive sample construction, i.e., the original positive pairs are well aligned in static scenes but differ in motions. Through our proposed cross-video motion alignment, the static bias issue is solved to some extent. It brings nearly 3% gains on both UCF-101 and HMDB-51 datasets compared to the baseline. In default, we discard the search on the static frame $\tilde{f}(s)$ and only maintain the extractor and queue of frame difference $\tilde{f}(d)$.

$\mathcal{L}_{ac}$	f_s	f_d	UCF-101	HMDB-51
-	-	-	61.1	34.4
✓			62.7	36.1
✓	✓		63.2	36.9
✓		✓	63.6	37.1
✓	✓	✓	64.3	37.7

Table 7: Ablation study on feature activation maps. We compare the results without cross-video retrieval.



(a) Activation maps A_s of static frame feature.



(b) Activation maps A_d of frame difference feature.

Figure 3: Activation maps of static frame and frame difference feature. We observe that A_s and A_d respectively attends to representative backgrounds and dynamic motions.

Besides the utilization of cross-video correspondence search, we compare the number of similar pairs and the sample distribution in our corrected motion positive samples. In Table 6, the first line denotes the baseline while the second line means only one sample is taken as positive pair. ‘Uniform’ indicates all pairs are equally important and the final representations are mean averaged. ‘Prior’ means the probability follows pair-wise similarity as Eq. 10. We can see that ‘Prior’ outperforms ‘Uniform’ consistently. It validates that the frame difference feature serves as a reliable reference for similar motion pattern retrieval. Interestingly, 5 positive samples achieve the highest accuracy while 10 get the worst. Theoretically, more abundant positive pairs lead to more accurate estimation, and there should be approximately $2048/101 \approx 20$ samples per class in the queue. However, due to the clip sampling, there are much fewer truly aligned motion patterns as illustrated in Fig. 1. Hence, the performance drops when we retrieve the ten nearest samples to constitute the positive pairs.

Feature Activation Maps. We detail the effect of using feature activation maps in Table 7. The first line denotes only decoupling in data inputs. It is clear that the activation alignment constraint $\mathcal{L}_{ac}$ is effective, and the static-dynamic feature decoupling f_s and f_d further improves performance. It is coincident with our motivation that the activation maps provide more concrete spatio-temporal reference to capture both static and dynamic characteristics and mitigate possible static bias.

4.4 Qualitative Analysis

To better understand how the activation maps work, we visualize the class-agnostic activation maps of the static frame and frame

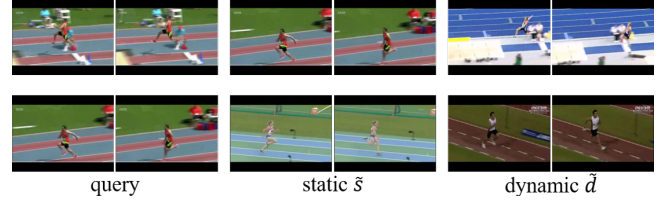


Figure 4: Cross-video search on static and dynamic features. We use two frames to present a clip, and show the Top-1 similar pair for both static scene s and dynamic motion d .

difference features in Fig. 3. The static frame features A_s attend to representative background areas like the ping pong table and baseball field. In contrast, the frame difference features A_d focus on the moving actors. Accordingly, A_s and A_d provide complementary views to understand a video and jointly contributes to unbiased video understanding.

Moreover, we also provide two typical examples of our cross-video retrieval results in Fig. 4. For better illustration, two queries are two views sampled from the same video. Through our cross-video search mechanism, we can figure out the exactly aligned clip with the query, especially in dynamic characters. For example, we match the query of the first row with another jumping moment as a motion pattern positive pair. And we find another running motion sample across the videos in the second row. Moreover, the static background of the motion pair appears totally different. It verifies that our cross-video search can reduce the background shortcut via the introduction of backgrounds from other videos. On the contrary, there was little difference between query and static similar pair $\tilde{s}$ in both motion and background content, which also echos with the quantitative results in Table 5 that the dynamic $\tilde{d}$ makes a great difference but the static $\tilde{s}$ has a minor effect.

5 CONCLUSION

In this paper, we propose a novel dual contrastive formulation to eliminate the static scene bias in spatio-temporal representation learning. We decouple the input RGB video sequence into the static frame and frame difference, then respectively minimize the decoupled loss term to guarantee the comprehensiveness of the RGB feature. We further utilize the static and dynamic activation maps as a concrete referrer to filter out redundancy that potentially interferes with learning. Through the experiments, we validate the effectiveness of dual contrastive learning formulation that simultaneously encode desired static and dynamic characteristics.

While our work shows some promising results, the state-of-the-art self-supervised performances [12] of UCF101 and HMDB51 are much higher than ours. But we believe our method can be further boosted through a huger model backbone with greater resolution input. We leave it as the feature work.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61932022, Grant 61971285, Grant 61720106001, and in part by the Program of Shanghai Science and Technology Innovation Project under Grant 20511100100.

REFERENCES

- [1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2019. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667* (2019).
- [2] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. 2020. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10451–10459.
- [3] Nadine Behrmann, Jurgen Gall, and Mehdi Noroozi. 2021. Unsupervised Video Representation Learning by Bidirectional Feature Prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1670–1679.
- [4] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. 2020. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9922–9931.
- [5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [6] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Minghui Tan, and Chuang Gan. 2021. Rspnet: Relative speed perception for unsupervised video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 1.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [8] Jinwoo Choi, Chen Gao, C. E. Joseph Messou, and Jia-Bin Huang. 2019. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. In *NeurIPS*.
- [9] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. 2022. Motion-Aware Contrastive Video Representation Learning via Foreground-Background Merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9716–9726.
- [10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*. 1422–1430.
- [11] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *arXiv preprint arXiv:2104.14548* (2021).
- [12] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. 2021. A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3299–3309.
- [13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* (2018).
- [14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhao-han Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733* (2020).
- [15] Tengda Han, Weidi Xie, and Andrew Zisserman. 2019. Video representation learning by dense predictive coding. In *Proceedings of the IEEE international conference on computer vision Workshops*. 0–0.
- [16] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Memory-augmented dense predictive coding for video representation learning. In *Proceedings of the European conference on computer vision*. Springer, 312–329.
- [17] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709* (2020).
- [18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [20] Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka. 2016. Human action recognition without human. In *European Conference on Computer Vision*. Springer, 11–17.
- [21] Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Minghui Tan, and Errui Ding. 2021. ASCNet: Self-supervised Video Representation Learning with Appearance-Speed Consistency. *arXiv preprint arXiv:2106.02342* (2021).
- [22] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. 2021. Self-supervised Video Representation Learning by Context and Motion Decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13886–13895.
- [23] Allan Jabri, Andrew Owens, and Alexei A Efros. 2020. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613* (2020).
- [24] Simon Jenni, Givi Meishvili, and Paolo Favaro. 2020. Video representation learning by recognizing temporal transformations. In *Proceedings of the European conference on computer vision*. Springer, 425–442.
- [25] Longlong Jing, Xiaodong Yang, Jingren Liu, and Yingli Tian. 2018. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387* (2018).
- [26] Dahun Kim, Donghyeon Cho, and In So Kweon. 2019. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8545–8552.
- [27] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. 2018. Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 793–802.
- [28] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. 2020. Cycle-contrast for self-supervised video representation learning. *arXiv preprint arXiv:2010.14810* (2020).
- [29] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. 2021. Mean Shift for Self-Supervised Learning. *arXiv preprint arXiv:2105.07269* (2021).
- [30] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Soren Schwertfeger, Cyrill Stachniss, and Mu Li. 2021. Video Contrastive Learning with Global Context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3195–3204.
- [31] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*. IEEE, 2556–2563.
- [32] Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. 2021. Motion-Focused Contrastive Learning of Video Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2105–2114.
- [33] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. 2019. Joint-task self-supervised learning for temporal correspondence. *arXiv preprint arXiv:1909.11895* (2019).
- [34] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. Resound: Towards action recognition without representation bias. In *Proceedings of the European conference on computer vision*. 513–528.
- [35] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. 2020. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11701–11708.
- [36] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. 2017. Unsupervised learning of long-term motion dynamics for videos. In *Proceedings of the IEEE international conference on computer vision*. 2203–2212.
- [37] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6707–6717.
- [38] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*. Springer, 527–544.
- [39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [40] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11205–11214.
- [41] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiya Lin. 2021. Enhancing Self-supervised Video Representation Learning via Multi-level Feature Optimization. *arXiv preprint arXiv:2108.02183* (2021).
- [42] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2020. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800* (2020).
- [43] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. 2021. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1255–1265.
- [44] Khuram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- [45] Li Tao, Xueting Wang, and Toshihiko Yamasaki. 2020. Self-supervised video representation learning using inter-intra contrastive framework. In *ACM MM*. 2193–2201.
- [46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. Springer, 776–794.
- [47] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.
- [48] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. 2017. Decomposing motion and content for natural video sequence prediction. *arXiv*

- preprint arXiv:1706.08033* (2017).
- [49] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE international conference on computer vision*. 98–106.
 - [50] Jinpeng Wang et al. 2021. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *AAAI*21.
 - [51] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. 2021. Removing the Background by Adding the Background: Towards Background Robust Self-supervised Video Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11804–11813.
 - [52] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Wei Liu, and Yun-hui Liu. 2020. Self-supervised Video Representation Learning by Uncovering Spatio-temporal Statistics. *arXiv preprint arXiv:2008.13426* (2020).
 - [53] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. 2020. Self-supervised video representation learning by pace prediction. In *Proceedings of the European conference on computer vision*.
 - [54] Xiaolong Wang, Allan Jabri, and Alexei A Efros. 2019. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2566–2576.
 - [55] Yang Wang and Minh Hoai. 2018. Pulling actions out of context: Explicit separation for effective combination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7044–7053.
 - [56] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3733–3742.
 - [57] Fanyu Xiao, Joseph Tighe, and Davide Modolo. 2021. MoDist: Motion Distillation for Self-supervised Video Representation Learning. *arXiv preprint arXiv:2106.09703* (2021).
 - [58] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision*. 305–321.
 - [59] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. 2019. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10334–10343.
 - [60] Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. 2021. Bag of Instances Aggregation Boosts Self-supervised Distillation. In *International Conference on Learning Representations*.
 - [61] Haohang Xu, Hongkai Xiong, and Guo-Jun Qi. 2021. K-Shot Contrastive Learning of Visual Features with Multiple Instance Augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
 - [62] Haohang Xu, Xiaopeng Zhang, Hao Li, Lingxi Xie, Wenrui Dai, Hongkai Xiong, and Qi Tian. 2022. Seed the views: Hierarchical semantic alignment for contrastive representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
 - [63] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. 2020. Seco: Exploring sequence supervision for unsupervised representation learning. *arXiv preprint arXiv:2008.00975* (2020).
 - [64] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. 2020. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6548–6557.
 - [65] Lin Zhang, Qi She, Zhengyang Shen, and Changhu Wang. 2021. How Incomplete is Contrastive Learning? An Inter-intra Variant Dual Representation Method for Self-supervised Video Recognition. *arXiv preprint arXiv:2107.01194* (2021).
 - [66] Zehua Zhang and David Crandall. 2022. Hierarchically decoupled spatial-temporal contrast for self-supervised video representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3235–3245.
 - [67] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. 2020. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606* (2020).
 - [68] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE international conference on computer vision*. 2921–2929.