Computer Vision and Image Understanding 219 (2022) 103406

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

TCLR: Temporal contrastive learning for video representation

Ishan Dave *, Rohit Gupta, Mamshad Nayeem Rizve, Mubarak Shah

Center for Research in Computer Vision, University of Central Florida, Orlando 32816, FL, United States of America

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
68T45
68T30
68T07

Keywords:
Self-Supervised Learning
Action Recognition
Video Representation

ABSTRACT

Contrastive learning has nearly closed the gap between supervised and self-supervised learning of image representations, and has also been explored for videos. However, prior work on contrastive learning for video data has not explored the effect of explicitly encouraging the features to be distinct across the temporal dimension. We develop a new temporal contrastive learning framework consisting of two novel losses to improve upon existing contrastive self-supervised video representation learning methods. The local–local temporal contrastive loss adds the task of discriminating between non-overlapping clips from the same video, whereas the global–local temporal contrastive aims to discriminate between timesteps of the feature map of an input clip in order to increase the temporal diversity of the learned features. Our proposed temporal contrastive learning framework achieves significant improvement over the state-of-the-art results in various downstream video understanding tasks such as action recognition, limited-label action classification, and nearest-neighbor video retrieval on multiple video datasets and backbones. We also demonstrate significant improvement in fine-grained action classification for visually similar classes. With the commonly used 3D ResNet-18 architecture with UCF101 pretraining, we achieve 82.4% (+5.1% increase over the previous best) top-1 accuracy on UCF101 and 52.9% (+5.4% increase) on HMDB51 action classification, and 56.2% (+11.7% increase) Top-1 Recall on UCF101 nearest neighbor video retrieval. Code released at https://github.com/DAVEISHAN/TCLR.

1. Introduction

Large-scale labeled datasets such as Kinetics (Carreira and Zisserman, 2017), LSHVU (Diba et al., 2020) etc have been crucial for recent advances in video understanding tasks. Since training a video encoder using existing supervised learning approaches is label-inefficient (Kataoka et al., 2020), annotated video data is required at a large scale. This costs enormous human effort and time, much more so than annotating images. At the same time, a tremendous amount of unlabeled video data is easily available on the internet. Research in self-supervised video representation learning can unlock the corpus of readily available unlabeled video data and unshackle progress in video understanding.

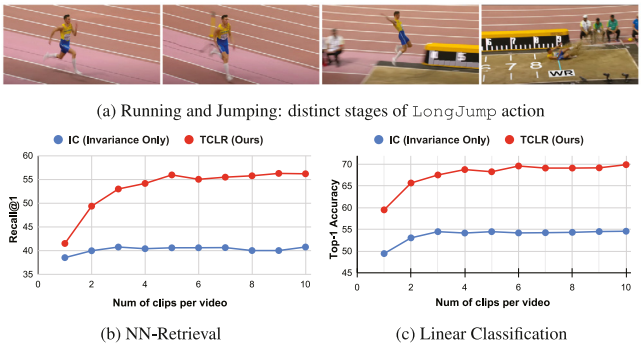
Recently, Contrastive Self-supervised Learning (CSL) based methods (Chen et al., 2020; He et al., 2020; Caron et al., 2020) have demonstrated the ability to learn powerful *image representations* in a self-supervised manner, and have narrowed down the performance gap between *unsupervised* and *supervised* representation learning on various image understanding downstream tasks.

A simple yet effective extension of CSL to the video domain can be obtained by using the InfoNCE instance discrimination objective, where the model learns to distinguish clips of a given video from the clips of other videos in the dataset (see Fig. 2a). Unlike images, videos have both time-invariant and the temporally varying properties. For example, in a `LongJump` video from UCF101 (See Fig. 1), running and jumping represent two very different stages of the action. Usually, video understanding models utilize temporally varying features by aggregating along the temporal dimension to obtain a video level prediction. While the significant success can be achieved on many video understanding tasks by only modeling the temporally invariant properties, it maybe possible that the temporally varying properties can also play an important role in further improvements on these tasks. Whether video representations should be invariant or distinct along the temporal dimension is an open question in the literature. Instance contrastive pre-training, however, encourages the model to learn similar features to represent temporally distant clips from the video, i.e. it enforces temporal invariance on the features. While instance level contrastive learning lies on one end of the spectrum, some recent works have tried to relax the invariance constraint through various means, such as, using a weighted temporal sampling strategy to avoid enforcing invariance between temporally distant clips (Qian et al., 2021a), cross-modal mining of positive samples from across video instances (Han et al., 2020b) or adding additional pretext tasks that require learning temporal features (Wang et al., 2020; Shao et al., 2021; Yao et al., 2021; Bai et al., 2020).

We take a different approach by explicitly encouraging the learning of temporally distinct video representations. The challenge with video

* Corresponding author.
E-mail address: ishandave@knights.ucf.edu (I. Dave).

https://doi.org/10.1016/j.cviu.2022.103406
Received 10 August 2021; Received in revised form 7 January 2022; Accepted 5 March 2022
Available online 16 March 2022



(a) Running and Jumping: distinct stages of LongJump action

(b) NN-Retrieval (c) Linear Classification

Fig. 1. Videos from standard action recognition datasets often have distinct temporal stages. For example, in figure **(a)** we can see the two distinct stages (Running and Jumping) of the LongJump action. Typically predictions across multiple short clips are aggregated, as a single short clip may not capture both stages of the action. We show the comparison of vanilla instance discrimination based contrastive (IC) self-supervision and our proposed TCLR method on **(b)** Nearest neighbor retrieval and **(c)** Linear classification tasks. We find that IC trained models do not benefit much from using multiple clips during evaluation. This is a result of IC imposing within instance temporal invariance. This motivates our proposed TCLR pre-training, which explicitly encourages learning distinct features across time.

classification is modeling variable length videos with a fixed number of parameters. 3D CNNs tackle this challenge by temporal aggregation of features across two levels: averaging across distinct fixed length temporal segments of a video (clips) and also temporal pooling across the feature map of each clip. Based on this observation, we propose two different temporal contrastive losses in order to learn temporally distinct features across the video: one which acts across clips of the same video, and another which acts across the timesteps of the feature map of the same clip. Combined with the vanilla instance contrastive loss, these novel losses result in an increase in the temporal diversity of the learned features, and better accuracy on downstream tasks.

Our first proposed loss is the *local–local temporal contrastive loss* (Fig. 2b), which ensures that temporally non-overlapping clips from the same video are mapped to distinct representations. This loss treats randomly augmented versions of the same clip as positive pairs to be brought together, and other non-overlapping clips from the same video as negative matches to be pushed away. While the local–local loss ensures that distinct clips have distinct representations, in order to encourage temporal variation *within each clip*, we introduce a second temporal contrastive loss, the *global–local temporal contrastive loss* (Fig. 2c). This loss constrains the *timesteps* of the feature map of a long *"global"* video clip to match the representations of the temporally aligned shorter *"local"* video clips.

Our complete framework is called *Temporal Contrastive Learning of video Representations* (henceforth referred to as *TCLR*). TCLR retains the ability of representations to successfully discriminate between video instances due to its instance contrastive loss. In addition, TCLR attempts to capture the *within-instance* temporal variation. Through extensive experiments on various downstream video understanding tasks, we demonstrate that both of our proposed Temporal Contrastive losses contribute to the learning of powerful video representations, and provide significant improvements.

The original contributions of this work can be summarized as below:

- TCLR is the first contrastive learning framework to explicitly enforce *within* instance temporal feature variation for video understanding tasks.
- Novel *local–local* and *global–local temporal contrastive losses*, which when combined with the standard instance contrastive loss significantly outperform the state-of-the-art on various downstream video understanding tasks like action recognition, nearest neighbor video retrieval and action classification with limited labeled data, while using 3 different 3D CNN architectures and 2 datasets (UCF101 & HMDB51).
- We propose the use of the challenging Diving48 fine-grained action classification task for evaluating the quality of learned video representations.

2. Related work

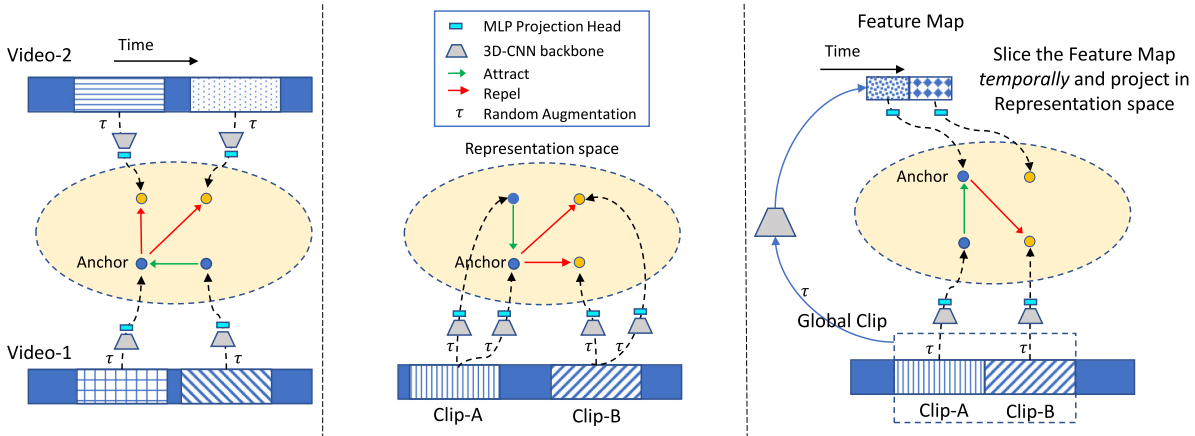
Recent approaches for self-supervised video representation learning can be categorized into two major groups based on the self-supervised learning objective: (1) Pretext task based methods, and (2) Contrastive Learning based methods.

Pretext task based approaches: Various pretext tasks have been devised for self-supervised video representation learning based on learning the correct temporal order of the data: verifying correct frame order (Misra et al., 2016), identifying the correctly ordered tuple from a set of shuffled orderings (Fernando et al., 2017; Suzuki et al., 2018), sorting frame order (Lee et al., 2017), and predicting clip order (Xu et al., 2019). Some methods extend existing pretext tasks from the image domain to video domain, for example, solving spatio-temporal jigsaw puzzles (Ahsan et al., 2019; Kim et al., 2019; Huo et al., 2021) and identifying the rotation of transformed video clips (Jing et al., 2018). Many recent works rely on predicting video properties like playback rate of the video (Cho et al., 2021; Yao et al., 2020a; Wang et al., 2020; Shao et al., 2021), temporal transformation that has been applied from a given set (Jenni et al., 2020; Jenni and Jin, 2021), speediness of moving objects (Benaim et al., 2020), and motion and appearance statistics of the video (Wang et al., 2019, 2021b).

Contrastive Self-supervised Learning (CSL) based approaches: Following the success of contrastive learning approaches of self-supervised image representation learning such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020), there have been many extensions of contrastive learning to the video domain. For instance, various video CSL methods (Wang et al., 2020; Pan et al., 2021; Bai et al., 2020; Qian et al., 2021a; Tao et al., 2020; Yao et al., 2021; Tokmakov et al., 2020; Shao et al., 2021; Yang et al., 2020; Feichtenhofer et al., 2021) leverage *Instance level Discrimination* objectives, and build their method upon them, where clips from the same video are treated as positives and clips from the different videos as negatives. CVRL (Qian et al., 2021a) studies the importance of temporal augmentation and develops a temporal sampler to avoid enforcing excessive temporal invariance in learning video representation. VideoMoCo (Pan et al., 2021) improves image-based MoCo framework for video representation by encouraging temporal robustness of the encoder and modeling temporal decay of the keys. VTHCL (Yang et al., 2020) employs SlowFast architecture (Feichtenhofer et al., 2019) and uses contrastive loss with the slow and fast pathway representations as the positive pair. VIE (Zhuang et al., 2020) is proposed as a deep neural embedding-based method to learn video representation in an unsupervised manner, by combining both static image representation from 2D CNN and dynamic motion representation from 3D CNN. Generative contrastive learning-based approaches such as predicting the dense representation of the next video block (Han et al., 2019, 2020a), or Contrastive Predictive Coding (CPC) (Oord et al., 2018) for videos (Lorre et al., 2020) have also been studied in the literature.

AMDIM (Bachman et al., 2019) is another CSL approach for image representation learning, where a local view (spatial slice of the feature map taken from an intermediate layer) and a global view (full feature map) of differently augmented versions of the same image are considered as a positive pair, and global views of other images form negative pair of the contrastive loss. The method is adapted for the video domain Devon Hjelm and Bachman (2020), Xue et al. (2020) by generating local views from the spatio-temporal features. Unlike this class of methods, which try to maximize agreement across features from different levels of the encoder, our Global–Local loss tries to learn distinct features across temporal slices of the feature map instead.

Some recent works combine pretext tasks along with contrastive learning in a multi-task setting to learn temporally varying features in



(a) Instance Contrastive Loss ($\mathcal{L}_{IC}$) *Instance discrimination* by maximizing agreement b/w clips which belong to the *same video*

(b) Local-Local Temporal Contrastive Loss ($\mathcal{L}_{LL}$) Learning distinct features for *non-overlapping* local clips from the same video by maximizing agreement b/w differently augmented clips which belong to the *same timestamp*, while repelling negatives from distant timestamps of the same video instance

(c) Global-Local Temporal Contrastive Loss ($\mathcal{L}_{GL}$) Learning distinct temporal features for the global clip by maximizing agreement b/w the *temporal feature slices* of the global clip and the representations of the local clips (Clip-A or B) from the *same timestamp*

Fig. 2. The proposed temporal contrastive learning framework (TCLR) for learning temporally distinct video representations consists of three different losses.

the video representation. For example, using video clips with different playback rates as positive pairs for contrastive loss along with predicting the playback rate (Wang et al., 2020), or temporal transforms (Bai et al., 2020). Other works propose frame-based contrastive learning, along with existing pretext tasks of frame rotation prediction (Knights et al., 2021) and frame-tuple order verification (Yao et al., 2021). Unlike these works, TCLR takes a different approach by adding explicit temporal contrastive losses that encourage temporal diversity in the learned features, instead of utilizing a pretext task for this purpose.

Some works which try to capture intra-video variance using optical flow, but are nevertheless interesting to compare with. IIC (Tao et al., 2020) uses intra-instance negatives, but it relies on frame repeating and shuffling to generate these “hard" negatives, and does not focus on learning distinct features across the temporal axis. DSM (Wang et al., 2021) tries to decouple scene and motion features by an intra-instance triplet loss, which uses negatives generated by optical flow scaling and spatial warping. Some recent works use extra supervisory signals in addition to the RGB video data to learn video representation in a self-supervised manner. However, these methods either require additional cross-modal data (e.g. text narration (Miech et al., 2020a), audio (Afouras et al., 2020)) or expensive and time-consuming computation of hand-crafted visual priors (e.g. optical flow (Tao et al., 2020; Sun et al., 2019; Wei et al., 2018; Tian et al., 2020; Han et al., 2020b) or dense trajectories (Tokmakov et al., 2020)). In this work we focus only on learning from RGB data without using any auxiliary data from any extra modality or additionally computed visual priors.

3. Method

The key idea in our proposed framework is to learn two levels of contrastive discrimination: instance discrimination using the *instance contrastive loss* and within-instance temporal level discrimination using our novel temporal contrastive losses. The two different temporal contrastive losses which are applied within the same video instance: *Local–Local Loss* and *Global–Local loss*. Each of these losses is explained in the following sections.

3.1. Instance contrastive loss

We leverage the idea of *instance discrimination* using InfoNCE (Gutmann and Hyvärinen, 2010) based contrastive loss for learning video representations. In the video domain, in addition to leveraging image-based spatial augmentations, temporal augmentations can also be applied to generate different transformed versions of a particular instance. For a video instance, we extract various clips (starting from different timestamps and/or having different frame sampling rates). We consider a randomly sampled mini-batch of size N_B from different video instances, and from each instance we extract a pair of clips from random timesteps resulting in a total of 2N clips. The extracted clips are augmented using standard stochastic appearance and geometric transformations.[1] Each of the transformed clips is then passed through a 3D-CNN based video encoder which is followed by a non-linear projection head (multi-layer perceptron) to project the encoded features on the representation space. Hence, for each video-instance i we get two clip representations (G_i, G_i'). The instance contrastive loss is defined as follows:

$$\mathcal{L}_{IC}^{i} = -\log \frac{h\left(G_i, G_i'\right)}{\sum_{j=1}^{N_B}[\mathbb{1}_{[j\neq i]}h(G_i, G_j) + h(G_i, G_j')]}, \tag{1}$$

where, $h(u, v) = \exp\left(u^T v/(\|u\|\|v\|\theta)\right)$ is used to compute the similarity between u and v vectors with an adjustable parameter temperature, θ. $\mathbb{1}_{[j\neq i]} \in \{0, 1\}$ is an indicator function which equals 1 iff $j \neq i$.

3.2. Temporal contrastive losses

For self-supervised training using the instance contrastive loss of Eq. (1), the model is presented with multiple clips cropped from random spatio-temporal locations within a single video as positive matches. This encourages the model to become invariant to the inherent variation present within an instance. In order to enable contrastive learning to represent within instance temporal variation, we introduce two novel temporal contrastive losses: *local–local temporal contrastive loss* and *global–local temporal contrastive loss*.

3.2.1. Local–local temporal contrastive loss

For this loss, we treat non-overlapping clips sampled from different temporal segments of the same video instance as negative pairs, and randomly transformed versions of the same clip as a positive pair.

[1] More details about the augmentations are available in Section 4 and Section C of the supplementary material.

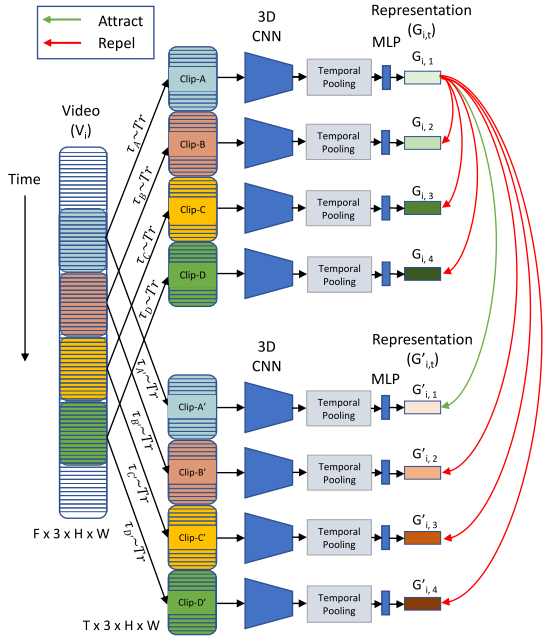


Fig. 3. Local–Local Temporal Contrastive Loss is applied to representations of non-overlapping clips extracted from same video instance i. For the clip starting at timestep p, two randomly transformed versions are generated and their representations $G_{i,p}$ and $G'_{i,p}$ serve as the positive pair for the loss, whereas the other non-overlapping clips along with the anchor, $G_{i,p}$, form the negative pairs. $p = 1$ serves as anchor, further details in Section 3.2.1. τ_A, $\tau_{A'}$, …, $\tau_{D'}$ are random set of augmentation sampled from universal set Tr.

The local–local loss is defined by Eq. (2) and illustrated in Fig. 3. A given video instance i is divided into N_{clips} non-overlapping clips. For the anchor clip starting at timestep p, its representation $G_{i,p}$, and the representation of its transformed version form the positive pair $(G_{i,p}, G'_{i,p})$) for this loss; whereas the other $N_{clips} - 1$ clips from the same video instance (and their transformed versions) form the negative pairs. Hence, for every positive pair, the local–local contrastive loss has $2 \times N_{clips} - 2$ negative pairs as defined in the following loss:

$$\mathcal{L}^i_{LL} = -\sum_{p=1}^{N_{clips}} \log \frac{h\left(G_{i,p}, G'_{i,p}\right)}{\sum_{q=1}^{N_{clips}} [\mathbb{1}_{[q\neq p]} h(G_{i,p}, G_{i,q}) + h(G_{i,p}, G'_{i,q})]}. \tag{2}$$

The key difference between the Instance contrastive loss (Eq. (1)) and the proposed local–local Temporal contrastive loss (Eq. (2)) is that for the local–local loss the negatives come from the same video instance but from a different temporal segment (clips), whereas in Eq. (1), the negative pairs come from different video instances.

3.2.2. Global–local temporal contrastive loss

The feature map in the higher layers of 3D CNNs are capable of representing temporal variation in the input clip, which is temporally pooled before being used for classification, or projected in the representation space in the case of contrastive learning. The objective of our proposed global–local temporal contrastive loss is to explicitly encourage the model to learn feature maps that represent the temporal locality of the input clip across temporal dimension of the feature map.

This loss is illustrated in Fig. 4. The notion of *local* and *global* is used at two different levels: at the input clip level and the feature level. Clip-E is a global clip and Clip A-D are local clips contained within Clip-A. Features are referred to as *global* after the final pooling operation and a temporal slice of the feature map before the pooling operation is referred to as a *local* feature. In Fig. 4, $L_{i,1}$ is the local feature of the global Clip-A and $G_{i,1}$ is the global feature of the local Clip-B.

For a video instance i, divided into N_{clips} clips, the local clip k can either be represented by a global (pooled) representation $G_{i,k}$ or a local

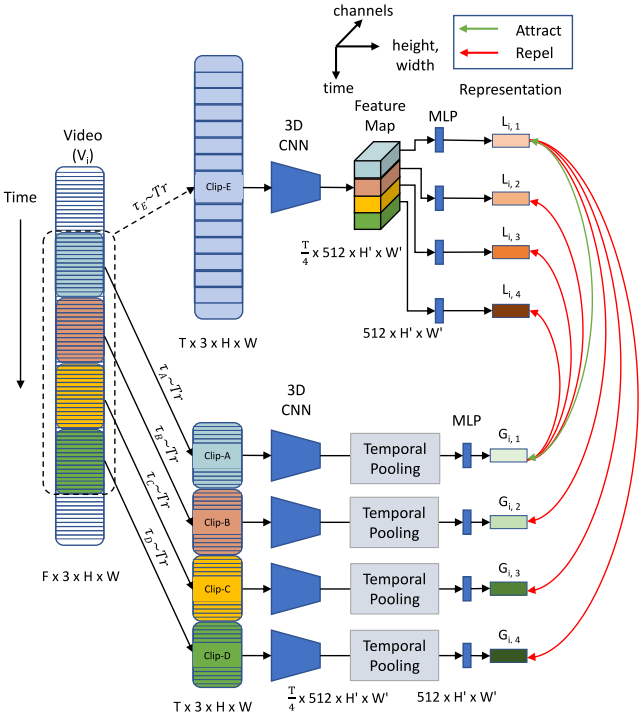


Fig. 4. Global–Local Temporal Contrastive Loss A global clip (Clip E) is extracted from a video instance and divided into 4 equal length local clips (Clips A through D). The global clip is temporally downsampled to have the same number of frames as each local clip. The local representations $L_{i,1}$ through $L_{i,4}$ from the global clip are obtained from the penultimate layer of the 3D-CNN (prior to temporal pooling). Global representations of the local clips, $G_{i,1}$ through $G_{i,4}$ are obtained from the CNN (after temporal pooling layer). This loss aims to maximize the similarity between the local representation of the global clip and the global representations of the corresponding local clip. Further details in Section 3.2.2.

representation $L_{i,k}$ of the corresponding timestep in the feature map of the global clip. This loss has two sets of reciprocal terms, with $G_{i,k}$ and $L_{i,k}$ serving as the anchor for each term. The negative pairs are supplied by matching the anchors with representations corresponding to other non-overlapping local clips. Note that similar to our local–local temporal contrastive loss we do not use negatives from other video instances for calculating this loss. The loss is defined by the following equations:

$$\mathcal{L}^i_{GL_k} = \log \frac{h\left(L_{i,k}, G_{i,k}\right)}{\sum_{q=1}^{N_{clips}} h(L_{i,k}, G_{i,q})} + \log \frac{h\left(G_{i,k}, L_{i,k}\right)}{\sum_{q=1}^{N_{clips}} h(G_{i,k}, L_{i,q})}, \tag{3}$$

$$\mathcal{L}^i_{GL} = -\sum_{k=1}^{N_{clips}} \mathcal{L}^i_{GL_k}. \tag{4}$$

4. Experiments

Datasets and Implementation: We use three action recognition datasets: UCF101 (Soomro et al., 2012), Kinetics400 (Carreira and Zisserman, 2017), and HMDB51 (Kuehne et al., 2011) for our experiments. We use the three most commonly used networks from the literature: 3D-ResNet-18 (R3D-18) (Hara et al., 2018), R(2+1)D-18 (Tran et al., 2018), and C3D (Tran et al., 2015) for our experiments. For non-linear projection head, we use a multi-layer perceptron with 1-hidden layer following experimental setting of Chen et al. (2020). We utilize 4 local clips per global clip for the global–local temporal contrastive loss. For all reported results, we utilize commonly used random augmentations including appearance-based transforms such as grayscale, channel dropping, and color jittering and geometry-based transforms like random scaling, random cropping, random cut-out and random horizontal-flip. Our results can be further improved by using more

complex augmentations like Gaussian blurring, shearing and rotation, however these are not used in the results reported in this paper. We provide results with more complex augmentation in Section D of the supplementary material. For self-supervised pretraining we use UCF101 training set (split-1) or Kinetics400 training set, without using any class labels. For all self-supervised pretraining, supervised finetuning and other downstream tasks, we use clips of 16 frames with a resolution of 112 × 112. More implementation details can be found in Section C of the supplementary material.

4.1. Evaluating self-supervised representations

We evaluate the learned video representation using different downstream video understanding tasks: (i) action recognition and (ii) nearest neighbor video retrieval on UCF101 and HMDB51 datasets, and (iii) limited label training on UCF101, following protocols from prior works (Han et al., 2020a). We also evaluate our learned representations on the challenging Diving-48 fine-grained action recognition task (Li et al., 2018); to the best of our knowledge TCLR is the first work that reports result on this challenging task. Our method is also employed in Knights (Dave et al., 2021) to get first place in ICCV-21 Action recognition challenge (Lengyel et al., 2022).

Action Recognition on UCF101 and HMDB51:

For the action recognition task on UCF101 and HMDB51, we first pretrain different video encoders in self-supervised manner on UCF101 or Kinetics400, and then perform supervised fine-tuning. In order to ensure fair comparison, we evaluate the method on the three most commonly used 3D CNN backbones in the prior works, while also listing details about the *input clip resolution* and *number of frames* used, as it is known to affect the results significantly (Tran et al., 2018; Patrick et al., 2021). Comparison results are shown in Table 1. Previous results based on multi-modal approaches that utilize text, audio etc are excluded (Patrick et al., 2021; Alwassel et al., 2020; Miech et al., 2020b). Results from prior works which do not utilize the three common architectures or use optical flow as input are presented in gray. We reproduce the results for CVRL (Qian et al., 2021a) using the R3D-18 model and 112 resolution, and carefully implement their temporal sampling and augmentation strategy. TCLR consistently outperforms the state-of-art by wide margins for all comparable combinations of backbone, pre-training dataset and fine-tuning dataset. The best prior results are reported by TaCo (Bai et al., 2020), which relies on learning temporal features using pretext tasks on top of instance discrimination. Our consistent improvement over TaCo suggests that using temporal contrastive losses results in better features than using existing temporal pre-text tasks in a multi-task setting.

Nearest Neighbor Video Retrieval:

We evaluate the learned representation by performing nearest neighbor retrieval after self-supervised pretraining on UCF101 videos and without any supervised finetuning. Videos from the test set are used as the query and the training set as the search gallery, following the protocol used in prior work (Han et al., 2020a). Results for retrieval are presented for both UCF101 and HMDB51 in Table 2. *TCLR outperforms previous state-of-the-art in UCF101 Top-1 Retrieval by 12% to 30% depending on the architecture*

Label Efficiency/ Finetuning with limited data: We evaluate our pretrained model for action recognition task on UCF101 (split-1) with limited labeled training data following the protocols from prior work (Han et al., 2020a; Jing et al., 2018; Gavrilyuk et al., 2021). Our method outperforms MotionFit (Gavrilyuk et al., 2021), MemDPC (Han et al., 2020a) and RotNet3D (Jing et al., 2018) in all settings of limited percentage of training data as shown in Fig. 5. This result in addition to NN results demonstrate that the learned representations from TCLR are significantly better than other recent works, *TCLR can achieve competitive performance to MemDPC with only 10% of the labeled data*.

Experiments on Diving-48 Dataset: This task presents some additional challenges over and above the standard action recognition task:

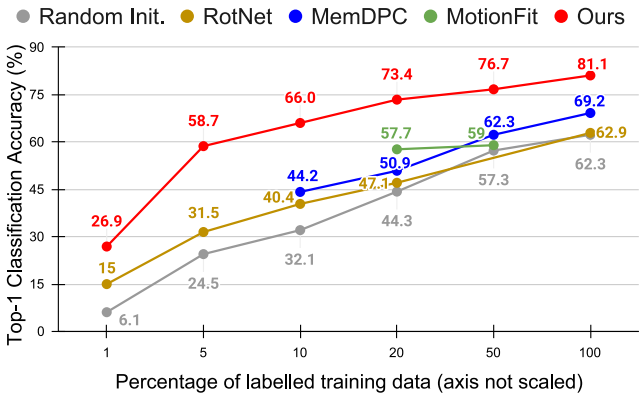


Fig. 5. Evaluating Label Efficiency using Limited Label Learning on UCF101 (split-1) action classification task.

action categories in Diving48 are defined by a combination of takeoff (dive groups), movements in flight (somersaults and/or twists), and entry (dive positions) stages. Two otherwise identical categories may only have fine grained differences limited to only one of the three stages. This makes Diving48 useful for evaluating the fine-grained representation capabilities of the model, which are not well tested by action recognition tasks on common benchmark datasets like UCF101 and HMDB51. Our proposed evaluation protocol consists of self-supervised pretraining followed by supervised finetuning on the Diving48-Train set. We adopt the 3D ResNet-18 architecture, with input resolution and clip length fixed at 112 × 112 and 16 frames, respectively.

Results are summarized in Table 3. TCLR pretraining on Diving48 without extra data outperforms random initialization and MiniKinetics (Xie et al., 2018) supervised pretraining. The within-instance temporal discrimination losses in TCLR help it outperform the instance contrastive loss. This is due to TCLR learning features to represent fine-grained differences between parts of diving video instances.

4.2. Ablation study

In order to study the impact of each contrastive loss used in TCLR, we test R3D-18 models pre-trained on UCF101 videos with a subset of the losses on each downstream task. The results for linear evaluation, full fine-tuning, transfer learning to HMDB51 and nearest neighbor retrieval are shown in Table 4. Addition of each temporal contrastive loss ($\mathcal{L}_{LL}$&$\mathcal{L}_{GL}$) leads to significant gains over instance contrastive and random initialization baselines, with the best results coming from combined use of all losses. We verify the correctness of our baselines by comparing them with similar results reported in prior work. Details can be found in Section E of the supplementary material. One interesting observation is that purely temporal contrastive learning, without instance discrimination, does not learn strong features directly (as can be seen from results on linear evaluation and NN-Retrieval), but it provides an useful initialization prior for supervised finetuning experiments.

4.3. Temporal diversity helps video understanding

To study the impact of temporal feature diversity directly, we utilize self-supervised pre-trained models only to avoid influence of supervised fine-tuning. As shown in Fig. 7, increasing from 1 clip to 10 clips per video, we observe that the pretraining strategies using temporal contrastive losses get significant performance gains (about *7–8% for each individual loss and 14.67% for TCLR*) with increasing number of clips. Instance Contrastive pretraining which enforces temporal invariance in learned features of video, does not see a similar improvement. It is also worth noting that each of the LL and GL losses help in learning different types of *temporal diversity* which results in TCLR having bigger

Table 1
Finetuning Results (average of 3 splits) for action classification on UCF101 and HMDB51. Self supervised pretraining was done on UCF101 (left) and Kinetics (right).† indicates models that utilize optical flow. ∗ indicates Kinetics-600 self-supervised pretraining. ‡ indicates ImageNet+Kinetics pre-training. Best and Second Best results are highlighted.

Method	Venue	Input Size	UCF101 Pre-Training		Kinetics400 Pre-Training	
			UCF101	HMDB51	UCF101	HMDB51
Backbone: R3D-18						
ST-Puzzle (Kim et al., 2019)	AAAI-19	16 × 112	–	–	65.8	33.7
STS (Wang et al., 2021b)	TPAMI-21	16 × 112	67.2	32.7	68.1	34.4
DPC (Han et al., 2019)	ICCVw-19	40 × 128	60.6	–	68.2	34.5
VCOP (Xu et al., 2019)	CVPR-20	16 × 112	64.9	29.5	–	–
Pace Pred (Wang et al., 2020)	ECCV-20	16 × 112	65.0	–	–	–
VCP (Luo et al., 2020)	AAAI-20	16 × 112	66.0	31.5	–	–
PRP (Yao et al., 2020a)	CVPR-20	16 × 112	66.5	29.7	–	–
Var. PSP (Cho et al., 2021)	Access-21	16 × 112	69.0	33.7	–	–
MemDPC (Han et al., 2020a)	ECCV-20	40 × 224	69.2	–	–	–
TCP (Lorre et al., 2020)	WACV-21	– × 224	64.8	34.7	70.5	41.1
VIE (Zhuang et al., 2020)	CVPR-20	16 × 112	–	–	72.3	44.8
UnsupIDT (Tokmakov et al., 2020)	ECCVw-20	16 × 112	–	–	73.0	41.6
CSJ (Huo et al., 2021)	–	16 × 224	70.4	36.0	76.2	46.7
BFP (Behrmann et al., 2021)	WACV-21	40 × 128	63.6	–	66.4	45.3
IIC (RGB) (Tao et al., 2020)	ACMMM-20	16 × 112	61.6	–	–	–
CVRL (Reproduced) (Qian et al., 2021a)	CVPR-21	16 × 112	75.77	44.6	–	–
SSTL (Shao et al., 2021)	–	16 × 112	–	–	79.1	49.7
VTHCL (Yang et al., 2020)	–	8 × 224	–	–	80.6	48.6
VideoMoCo (Pan et al., 2021)	CVPR-21	16 × 112	–	–	74.1	43.6
RSPNet (Chen et al., 2021)	AAAI-21	16 × 112	–	–	74.3	41.8
Temp Trans (Jenni et al., 2020)	ECCV-20	16 × 112	77.3	47.5	79.3*	49.8*
TaCo (Bai et al., 2020)	–	16 × 224	–	–	81.4	45.4
MFO (Qian et al., 2021b)	ICCV-21	16 × 112	–	–	79.1	47.6
TCLR		16 × 112	82.4	52.9	84.1	53.6
TCLR (Best Ablation)		16 × 112	83.9	53.5	85.4	55.4
Backbone: R(2+1)D-18						
VCP (Luo et al., 2020)	AAAI-20	16 × 112	66.3	32.2	–	–
PRP (Yao et al., 2020a)	CVPR-20	16 × 112	72.1	35.0	–	–
VCOP (Xu et al., 2019)	CVPR-20	16 × 112	72.4	30.9	–	–
Pace Pred (Wang et al., 2020)	ECCV-20	16 × 112	75.9	35.9	77.1	36.6
STS (Wang et al., 2021b)	TPAMI-21	16 × 112	73.6	34.1	77.8	40.5
VideoMoCo (Pan et al., 2021)	CVPR-21	16 × 112	–	–	78.7	49.2
VideoDIM (Devon Hjelm and Bachman, 2020)	–	32 × 128	–	–	79.7*	49.2*
RSPNet (Chen et al., 2021)	AAAI-21	16 × 112	–	–	81.1	44.6
Temp Trans (Jenni et al., 2020)	ECCV-20	16 × 112	81.6	46.4	–	–
TaCo (Bai et al., 2020)	–	16 × 224	–	–	81.8	46.0
TCLR		16 × 112	82.8	53.6	88.2	60.0
Backbone: C3D						
MA Stats-1 (Wang et al., 2019)	CVPR-19	16 × 112	58.8	32.6	61.2	33.4
Temp Trans (Jenni et al., 2020)	ECCV-20	16 × 112	68.3	38.4	69.9*	39.6*
PRP (Yao et al., 2020a)	CVPR-20	16 × 112	69.1	34.5	–	–
VCP (Luo et al., 2020)	AAAI-20	16 × 112	68.5	32.5	–	–
VCOP (Xu et al., 2019)	CVPR-20	16 × 112	65.6	28.4	–	–
Pace Pred (Wang et al., 2020)	ECCV-20	16 × 112	68.0	–	–	–
STS (Wang et al., 2021b)	TPAMI-21	16 × 112	69.3	34.2	71.8	37.8
Var. PSP (Cho et al., 2021)	Access-21	16 × 112	70.4	34.3	–	–
DSM (Wang et al., 2021)	AAAI-21	16 × 112	70.3	40.5	–	–
TCLR		16 × 112	76.1	48.6	–	–
Other Configurations						
CVRL (R3D-50) (Qian et al., 2021a)	CVPR-21	32 × 224	–	–	92.2	66.7
RSPNet (S3D-G) (Chen et al., 2021)	AAAI-21	64 × 224	–	–	93.7	64.7
CoCLR† (S3D-23) (Han et al., 2020b)	NeurIPS-20	16 × 112	87.3	58.7	90.6	62.9
SpeedNet (S3D-G) (Benaim et al., 2020)	CVPR-20	16 × 224	–	–	81.1	48.8
ρSimCLR (R50) (Feichtenhofer et al., 2021)	CVPR-21	8 × 224	-	-	85.6	-
SeCO (R50+TSN) (Yao et al., 2021)	AAAI-21	50 × 224	–	–	88.3‡	55.6‡

improvements relative to either of the temporal contrastive losses. Performance gains of a similar nature can also be observed in other downstream tasks as well, which are reported in Section F of the supplementary material.

4.4. Distinguishing confusing class pairs

To examine the ability of TCLR to distinguish confusing classes, we looked at the most confused action class pairs for UCF101 action recognition models trained from scratch. We observe that the these pairs mostly consist of fine-grained variants of action classes, for example the swimming actions `BreastStroke` and `FrontCrawl`. Some such pairs of classes are visualized in Fig. 6d. We can see that these classes are confusing because the corresponding frames are visually similar. In this study we considered a model without pretraining as a baseline, and tried to see the impact of instance contrastive and TCLR pretraining on it. We observe that despite a significant overall

Table 2
Nearest neighbor video retrieval results on UCF101 and HMDB51, after self-supervised pretraining on UCF101. * marks models pretrained on Kinetics-400. Best and second best results highlighted. Methods based on optical flow and audio modalities are excluded.

Method	Venue	UCF101				HMDB51			
		R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Backbone: R3D-18									
VCOP (Xu et al., 2019)	CVPR-20	14.1	30.3	40.4	51.1	7.6	22.9	34.4	48.8
VCP (Luo et al., 2020)	AAAI-20	18.6	33.6	42.5	53.5	7.6	24.4	36.6	53.6
Pace Pred (Wang et al., 2020)	ECCV-20	23.8	38.1	46.4	56.6	9.6	26.9	41.1	56.1
Var. PSP (Cho et al., 2021)	Access-21	24.6	41.9	51.3	62.7	10.3	26.6	38.8	51.6
Temp Trans (Jenni et al., 2020)	ECCV-20	26.1	48.5	59.1	69.6	–	–	–	–
STS* (Wang et al., 2021b)	TPAMI-21	38.3	59.9	68.9	77.2	18.0	37.2	50.7	64.8
SSTL* (Shao et al., 2021)	–	44.5	57.4	63.5	70.0	21.8	35.7	44.2	57.7
CSJ* (Huo et al., 2021)	–	21.5	40.5	53.2	64.9	–	–	–	–
MemDPC (Han et al., 2020a)	ECCV-20	20.2	40.4	52.4	64.7	7.7	25.7	40.6	57.7
RSPNet (Chen et al., 2021)	AAAI-21	41.1	59.4	68.4	77.8	–	–	–	–
MFO (Qian et al., 2021b)	ICCV-21	39.6	57.6	69.2	78.0	18.8	39.2	51.0	63.7
TCLR	–	56.2	72.2	79.0	85.3	22.8	45.4	57.8	73.1
Backbone: C3D									
VCOP (Xu et al., 2019)	CVPR-20	12.5	29.0	39.0	50.6	5.7	19.5	30.7	45.8
VCP (Luo et al., 2020)	AAAI-20	17.3	31.5	42.0	52.6	7.8	23.8	35.3	49.3
Pace Pred (Wang et al., 2020)	ECCV-20	31.9	49.7	59.2	68.9	12.5	32.2	45.4	61.0
DSM (Wang et al., 2020)	AAAI-21	16.8	33.4	43.4	54.6	8.2	25.9	38.1	52.0
STS* (Wang et al., 2021b)	TPAMI-21	39.1	59.2	68.8	77.6	16.4	36.9	49.9	64.9
RSPNet (Chen et al., 2021)	AAAI-21	36.0	56.7	66.5	76.3	–	–	–	–
TCLR	–	48.6	67.6	75.5	82.5	19.3	43.3	57.6	70.1
Backbone: R(2+1)D-18									
VCOP (Xu et al., 2019)	CVPR-20	10.7	25.9	35.4	47.3	7.4	22.6	34.4	48.5
VCP (Luo et al., 2020)	AAAI-20	19.9	33.7	42.0	50.5	6.7	21.3	32.7	49.2
Pace Pred (Wang et al., 2020)	ECCV-20	25.6	42.7	51.3	61.3	12.9	31.6	43.2	58.0
STS* (Wang et al., 2021b)	TPAMI-21	38.1	58.9	68.1	77.0	16.4	36.9	50.5	65.4
TCLR	–	56.9	72.2	79.0	84.6	24.1	45.8	58.3	75.3

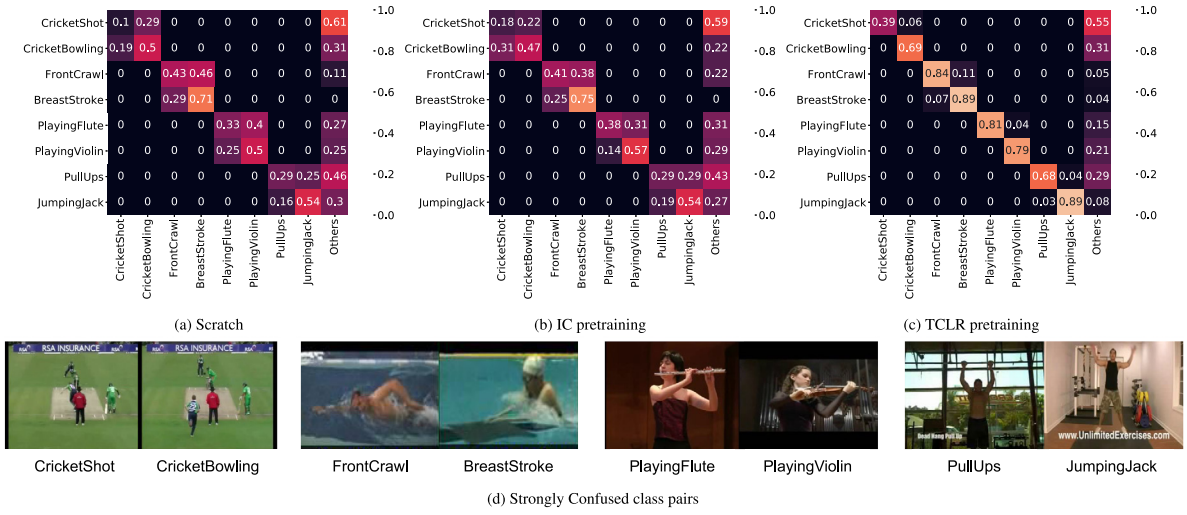


Fig. 6. Confusion matrices for 4 highly confused class-pairs from UCF101 classification models with **(a)** no pretraining, **(b)** IC pretraining, and **(c)** TCLR pretraining. **(d)** Classes illustrated with a sample frame. TCLR significantly improves over IC in distinguishing visually similar classes. .

Table 3
Diving48 fine-grained action classification results.

Pre-Training	Accuracy
None (Random Initialization)	13.4
MiniKinetics Supervised (Choi et al., 2019)	18.0
Instance Contrastive	15.8
VCOP (Xu et al., 2019)	14.7
CVRL (Qian et al., 2021a)	17.6
TCLR	22.9

improvement in accuracy, instance contrastive pre-training does not provide any significant gain in distinguishing these confused class pairs over the scratch baseline. On the other hand, TCLR pre-training helps remarkably with the confused classes. Average recall for these 8 classes is 42.5% for the scratch model, 44.9% for the IC model and 74.8% for the TCLR model. Since the classes are visually similar, distinguishing them requires learning the temporal variation in videos.

5. Conclusion

In this work, we propose two novel Temporal Contrastive losses to improve the quality of learned self-supervised video representations over standard instance discrimination contrastive learning. We provide extensive experimental evidence on three diverse datasets and obtain state-of-the-art results across various downstream video

Table 4
Ablation study of the **impact of temporal contrastive losses** on downstream tasks. Green indicates improvements over instance contrastive baseline.

Contrastive Losses			Classification Top1 Acc.			Retrieval
$\mathcal{L}_{IC}$	$\mathcal{L}_{LL}$	$\mathcal{L}_{GL}$	Linear Eval UCF101	Finetune UCF101	Transfer HMDB51	R@1 UCF101
Random Init.			17.15	62.39	26.95	8.21
✗	✓	✗	21.58	68.42	–	13.66
✗	✗	✓	20.61	70.19	–	12.83
✗	✓	✓	23.39	74.29	47.35	14.17
✓	✗	✗	54.58	71.31	38.32	40.76
✓	✓	✗	62.70+8%	77.70+6%	49.77+11%	51.10+10%
✓	✗	✓	64.55+10%	76.30+5%	47.87+10%	47.32+7%
✓	✓	✓	**69.91**+15%	**82.40**+11%	**52.80**+14%	**56.17**+15%

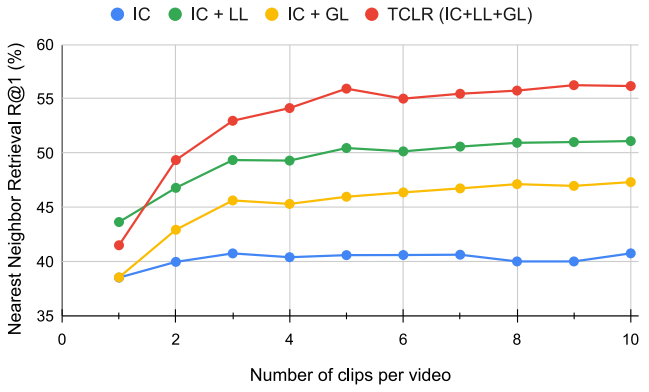


Fig. 7. Temporally distinct features learned by TCLR result in a significant improvement in NN-Retrieval on UCF101 (split-1) with increasing number of clips per video.

understanding tasks. The success of our approach underscores the benefits of contrastive learning beyond instance discrimination.

CRediT authorship contribution statement

Ishan Dave: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Rohit Gupta:** Methodology, Writing – review & editing, Visualization. **Mamshad Nayeem Rizve:** Writing – review & editing, Methodology, Validation. **Mubarak Shah:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Ishan Dave would like to acknowledge support from the Office of the Director of National Intelligence (ODNI), United States of America, Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2022.103406.

References

Afouras, T., Owens, A., Chung, J.S., Zisserman, A., 2020. Self-supervised learning of audio-visual objects from video. In: The European Conference on Computer Vision. ECCV.

Ahsan, U., Madhok, R., Essa, I., 2019. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In: 2019 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 179–189.

Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D., 2020. Self-supervised learning by cross-modal audio-video clustering. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., pp. 9758–9770.

Bachman, P., Hjelm, R.D., Buchwalter, W., 2019. Learning representations by maximizing mutual information across views. In: Advances in Neural Information Processing Systems. pp. 15535–15545.

Bai, Y., Fan, H., Misra, I., Venkatesh, G., Lu, Y., Zhou, Y., Yu, Q., Chandra, V., Yuille, A., 2020. Can temporal information help with contrastive self-supervised learning? arXiv preprint arXiv:2011.13046.

Behrmann, N., Gall, J., Noroozi, M., 2021. Unsupervised video representation learning by bidirectional feature prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1670–1679.

Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T., 2020. SpeedNet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9922–9931.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., pp. 9912–9924.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308.

Chen, P., Huang, D., He, D., Long, X., Zeng, R., Wen, S., Tan, M., Gan, C., 2021. RSPNet: Relative speed perception for unsupervised video representation learning. In: The AAAI Conference on Artificial Intelligence. AAAI.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: ICML.

Cho, H., Kim, T., Chang, H.J., Hwang, W., 2021. Self-supervised visual learning by variable playback speeds prediction of a video. IEEE Access 9, 79562–79571. http://dx.doi.org/10.1109/ACCESS.2021.3084840.

Choi, J., Gao, C., Messou, J.C., Huang, J.-B., 2019. Why can't I dance in the mall? Learning to mitigate scene bias in action recognition. In: Advances in Neural Information Processing Systems. pp. 853–865.

Dave, I., Biyani, N., Clark, B., Gupta, R., Rawat, Y., Shah, M., 2021. "Knights": first place submission for vipriors21 action recognition challenge at iccv 2021. arXiv preprint arXiv:2110.07758.

Devon Hjelm, R., Bachman, P., 2020. Representation learning with video deep InfoMax. arXiv preprint arXiv:2007.13278.

Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L., 2020. Large scale holistic video understanding. In: European Conference on Computer Vision. Springer, pp. 593–610.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6202–6211.

Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K., 2021. A large-scale study on unsupervised spatiotemporal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3299–3309.

Fernando, B., Bilen, H., Gavves, E., Gould, S., 2017. Self-supervised video representation learning with odd-one-out networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3636–3645.

Gavrilyuk, K., Jain, M., Karmanov, I., Snoek, C.G., 2021. Motion-augmented self-training for video recognition at smaller scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10429–10438.

Gutmann, M., Hyvärinen, A., 2010. 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.

Han, T., Xie, W., Zisserman, A., 2019. Video representation learning by dense predictive coding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops.

Han, T., Xie, W., Zisserman, A., 2020a. Memory-augmented dense predictive coding for video representation learning. In: Computer Vision–ECCV 2020: 16th European Conference. Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, pp. 312–329.

Han, T., Xie, W., Zisserman, A., 2020b. Self-supervised co-training for video representation learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., pp. 5679–5690.

Hara, K., Kataoka, H., Satoh, Y., 2018. Towards good practice for action recognition with spatiotemporal 3D convolutions. In: 2018 24th International Conference on Pattern Recognition. ICPR, pp. 2516–2521.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

Huo, Y., Ding, M., Lu, H., Lu, Z., Xiang, T., Wen, J.-R., Huang, Z., Jiang, J., Zhang, S., Tang, M., Huang, S., Luo, P., 2021. Self-supervised video representation learning with constrained spatiotemporal jigsaw. URL https://openreview.net/forum?id=4AWko4A35ss.

Jenni, S., Jin, H., 2021. Time-equivariant contrastive video representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9970–9980.

Jenni, S., Meishvili, G., Favaro, P., 2020. Video representation learning by recognizing temporal transformations. In: The European Conference on Computer Vision. ECCV.

Jing, L., Yang, X., Liu, J., Tian, Y., 2018. Self-supervised spatiotemporal feature learning via video rotation prediction. arXiv preprint arXiv:1811.11387.

Kataoka, H., Wakamiya, T., Hara, K., Satoh, Y., 2020. Would mega-scale datasets further enhance spatiotemporal 3D cnns? arXiv preprint arXiv:2004.04968.

Kim, D., Cho, D., Kweon, I.S., 2019. Self-supervised video representation learning with space-time cubic puzzles. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33. pp. 8545–8552.

Knights, J., Harwood, B., Ward, D., Vanderkop, A., Mackenzie-Ross, O., Moghadam, P., 2021. Temporally coherent embeddings for self-supervised video representation learning. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 8914–8921.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision. ICCV.

Lee, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H., 2017. Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 667–676.

Lengyel, A., Bruintjes, R.-J., Rios, M.B., Kayhan, O.S., Zambrano, D., Tomen, N., van Gemert, J., 2022. Vipriors 2: visual inductive priors for data-efficient deep learning challenges. arXiv preprint arXiv:2201.08625.

Li, Y., Li, Y., Vasconcelos, N., 2018. Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 513–528.

Lorre, G., Rabarisoa, J., Orcesi, A., Ainouz, S., Canu, S., 2020. Temporal contrastive pretraining for video action recognition. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 662–670.

Luo, D., Liu, C., Zhou, Y., Yang, D., Ma, C., Ye, Q., Wang, W., 2020. Video cloze procedure for self-supervised spatio-temporal learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33. pp. 11701–11708.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A., 2020a. End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A., 2020b. End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889.

Misra, I., Zitnick, C.L., Hebert, M., 2016. Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. Springer, pp. 527–544.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W., 2021. 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.

Patrick, M., Asano, Y., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A., 2021. Multi-modal self-supervision from generalized data transformations. URL https://openreview.net/forum?id=mgVbI13p96.

Qian, R., Li, Y., Liu, H., See, J., Ding, S., Liu, X., Li, D., Lin, W., 2021b. Enhancing self-supervised video representation learning via multi-level feature optimization. In: Proceedings of the International Conference on Computer Vision. ICCV.

Qian, R., Meng, T., Gong, B., Yang, M.-H., Wang, H., Belongie, S., Cui, Y., 2021a. Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974.

Shao, H., Liu, Y., Li, H., 2021. Self-supervised temporal learning. URL https://openreview.net/forum?id=WEnXA3sAwV7.

Soomro, K., Zamir, A.R., Shah, M., 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

Sun, C., Baradel, F., Murphy, K., Schmid, C., 2019. Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743.

Suzuki, T., Itazuri, T., Hara, K., Kataoka, H., 2018. Learning spatiotemporal 3D convolution with video order self-supervision. In: Proceedings of the European Conference on Computer Vision. ECCV.

Tao, L., Wang, X., Yamasaki, T., 2020. Self-supervised video representation learning using inter-intra contrastive framework. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2193–2201.

Tian, Y., Che, Z., Bao, W., Zhai, G., Gao, Z., 2020. Self-supervised motion representation via scattering local motion cues. In: Computer Vision–ECCV 2020: 16th European Conference. Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, pp. 71–89.

Tokmakov, P., Hebert, M., Schmid, C., 2020. Unsupervised learning of video representations via dense trajectory clustering. In: European Conference on Computer Vision. Springer, pp. 404–421.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6450–6459.

Wang, J., Gao, Y., Li, K., Jiang, X., Guo, X., Ji, R., Sun, X., 2021. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In: The AAAI Conference on Artificial Intelligence. AAAI.

Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W., 2019. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4006–4015.

Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W., 2021b. Self-supervised video representation learning by uncovering spatio-temporal statistics. IEEE Trans. Pattern Anal. Mach. Intell..

Wang, J., Jiao, J., Liu, Y.-H., 2020. Self-supervised video representation learning by pace prediction. In: The European Conference on Computer Vision. ECCV.

Wei, D., Lim, J.J., Zisserman, A., Freeman, W.T., 2018. Learning and using the arrow of time. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8052–8060.

Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K., 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 305–321.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y., 2019. Self-supervised spatiotemporal learning via video clip order prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10334–10343.

Xue, F., Ji, H., Zhang, W., Cao, Y., 2020. Self-supervised video representation learning by maximizing mutual information. Signal Process., Image Commun. 88, 115967.

Yang, C., Xu, Y., Dai, B., Zhou, B., 2020. Video representation learning with visual tempo consistency. arXiv preprint arXiv:2006.15489.

Yao, Y., Liu, C., Luo, D., Zhou, Y., Ye, Q., 2020a. 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.

Yao, T., Zhang, Y., Qiu, Z., Pan, Y., Mei, T., 2021. Seco: exploring sequence supervision for unsupervised representation learning. In: AAAI. 2, p. 7.

Zhuang, C., She, T., Andonian, A., Mark, M.S., Yamins, D., 2020. Unsupervised learning from video with deep neural embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9563–9572.