Contrastive Spatio-Temporal Pretext Learning for Self-Supervised Video Representation

Yujia Zhang[1]*, Lai-Man Po[1], Xuyuan Xu[2], Mengyang Liu[2], Yexin Wang[2], Weifeng Ou[1], Yuzhi Zhao[1], Wing-Yin Yu[1]

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China
[2]AI Technology Center, OVB, Tencent, Shenzhen, China
{yzhang2383-c,weifengou2-c,yzzhao2-c,wingyinyu8-c}
@my.cityu.edu.hk, eelmpo@cityu.edu.hk, {xuyuanxu2,yexin.w}@gmail.com, myleonliu@tencent.com

Abstract

Spatio-temporal representation learning is critical for video self-supervised representation. Recent approaches mainly use contrastive learning and pretext tasks. However, these approaches learn representation by discriminating sampled instances via feature similarity in the latent space while ignoring the intermediate state of the learned representations, which limits the overall performance. In this work, taking into account the degree of similarity of sampled instances as the intermediate state, we propose a novel pretext task - spatio-temporal overlap rate (STOR) prediction. It stems from the observation that humans are capable of discriminating the overlap rates of videos in space and time. This task encourages the model to discriminate the STOR of two generated samples to learn the representations. Moreover, we employ a joint optimization combining pretext tasks with contrastive learning to further enhance the spatio-temporal representation learning. We also study the mutual influence of each component in the proposed scheme. Extensive experiments demonstrate that our proposed STOR task can favor both contrastive learning and pretext tasks. The joint optimization scheme can significantly improve the spatio-temporal representation in video understanding. The code is available at https://github.com/Katou2/CSTP.

Introduction

Convolutional Neural Networks (CNNs) have been proven to be successful in supervised video representation learning with numerous human-annotated labels (Carreira and Zisserman 2017; Feichtenhofer et al. 2019). Videos contain more complex spatio-temporal contents and a larger data volume. Billions of unlabeled videos emerge on the Internet every day, making supervised video analysis expensive and time-consuming. Thus, how to effectively learn video representations without annotations is an important yet challenging task. Among effective unsupervised learning methods, self-supervised learning has proven to be a promising methodology (Chen et al. 2020; He et al. 2020; Feichtenhofer et al. 2021).

Early video self-supervised learning approaches proposed proper tasks with automatically generated labels, thereby encouraging CNNs to learn the transferable features for downstream tasks without human-annotated labels (Fernando et al. 2017; Benaim et al. 2020; Behrmann, Gall, and Noroozi 2021). Recently, with the success of contrastive learning in self-supervised image classification, this method has been widely expended in video self-supervised learning (Chen et al. 2021; Alwassel et al. 2019; Sermanet et al. 2018).

However, there are obvious limitations in these works. Firstly, previous works only explored discriminating similar features from dissimilar ones while ignoring the intermediate state of learned representations such as the similarity degree, which limits the overall performance. Secondly, less effort has been put on the mutual influence of multiple pretext tasks and various contrastive learning schemes for spatio-temporal representation learning.

To address these problems, we propose a novel pretext task, i.e., Spatio-Temporal Overlap Rate (STOR) prediction to percept the similarity degree as the intermediate state to favor contrastive learning and propose a joint optimization framework of contrastive learning and multiple pretext tasks to further enhance the spatio-temporal representation learning. It is observed that given a set of two clips at a specific overlap rate, humans can discriminate the overlap rate when providing candidates (see Figure 1). We assume that humans can make it due to their favorable spatio-temporal representation ability. Built upon the observation, we believe that CNNs can learn video representations better by discriminating such overlap rates. The assumption is that CNNs can only succeed in such a spatio-temporal overlap rate reasoning task when it learns representative spatio-temporal features. To the best of our knowledge, this is the first work that attempts to capture the spatio-temporal degree of similarity between generated samples for self-supervised learning. Moreover, we propose a new and effective data augmentation method for the pretext task. This data augmentation method can generate samples with random overlapped spatio-temporal regions, while keeping the randomness of samples.

In order to study the mutual influence of contrastive learning and pretext tasks for better spatio-temporal representation learning, we comprehensively study the joint optimization scheme. Specifically, we study four popular contrastive learning frameworks, i.e., SimCLR (Chen et al. 2020), MoCo (He et al. 2020), BYOL (Grill et al. 2020) and SimSi-

*Corresponding author.

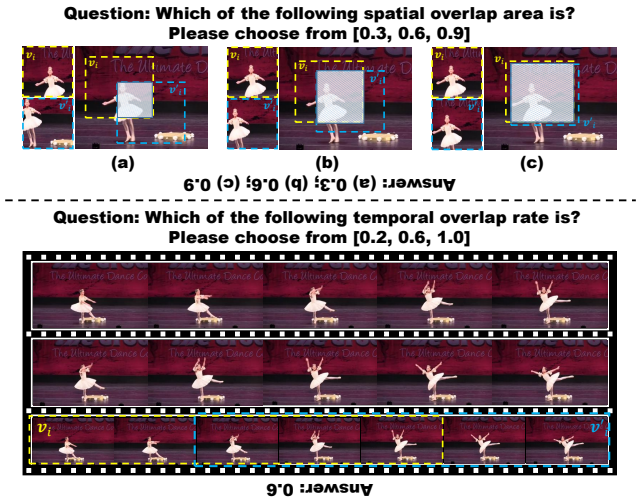


Figure 1: Illustration of the proposed overlap rate prediction in space and time. Given a video sample, the first clip v_i (Yellow) is randomly cropped in space and time, the second clip v'_i (Blue) is randomly cropped according to the size of bounding boxes and time duration of v_i at a random spatial and temporal overlap rate.

am (Chen and He 2021) in our joint optimization scheme. In terms of pretext tasks, playback rate prediction has been proven to be successful in video self-supervised learning (Wang, Jiao, and Liu 2020; Jenni, Meishvili, and Favaro 2020), but it tends to focus on motion pattern thus may not learn spatial pattern well (Chen et al. 2021). To overcome this problem, we combine rotation prediction task to further strengthen spatial appearance features. Finally, a joint optimization framework of STOR, contrastive learning, playback rate prediction and rotation prediction is proposed in this work, namely contrastive spatio-temporal pretext learning (CSTP).

Extensive experimental evaluations on two downstream video understanding tasks demonstrate the effectiveness of the proposed approach. Specifically, several architectures including C3D (Tran et al. 2015), R(2+1)D (Tran et al. 2018) and S3D (Xie et al. 2018) are presented and different weights of the joint learning are explored in this work. The experimental results verifies that the proposed STOR can well cooperate with different contrastive learning frameworks and other pretext tasks. The proposed joint learning framework CSTP outperforms state-of-the-art approaches in the two downstream video understanding tasks.

The main contributions of this work can be summarized as follows:

- Taking the degree of similarity of training samples into account, we propose a novel pretext task, i.e., spatio-temporal overlap rate prediction for video self-supervised learning. The pretext task can enhance spatio-temporal representation learning via discriminating the overlap regions of the training samples.
- We propose a joint optimization framework which combines contrastive learning with spatio-temporal pretext tasks. And we conduct comprehensive experiments to study the mutual influence of each component of the framework.
- Our method achieves state-of-the-art performance on two downstream tasks, action recognition and video retrieval, across two datasets, UCF-101 and HMDB-51. Ablation studies demonstrate the efficacy of the proposed STOR and the mutual influence of contrastive learning and pretext tasks.

Related Works

Self-supervised Learning in Images

Current image self-supervised methods can be briefly grouped into two types of paradigms, i.e., contrastive learning and pretext tasks. Early work (Dosovitskiy et al. 2015) studied self-supervised learning by setting each sample in the dataset as a category to train a classification model. However, the approach will be infeasible when the size of dataset is huge. To settle this problem, Wu et al. (Wu et al. 2018) employed a memory bank of previous features of samples to replace the classifier. He et al. (He et al. 2020) proposed a dynamic dictionary with a queue for memory bank and a momentum encoder to maintain the consistency of samples. Chen et al. (Chen et al. 2020) used a larger batch size to fully replace the memory bank. Grill et al. (Grill et al. 2020) proposed to use a MLP as feature predictor to replace the memory bank. Recently, Chen and He (Chen and He 2021) showed that using stop-gradient method, a Siamese architecture without memory bank and a MLP feature predictor can achieve state-of-the-art results. Also, self-supervised learning approaches based on pretext tasks have been extensively studied. Examples include recovering the input with appearance transformations, e.g., image colorization (Zhao et al. 2020; Su, Chu, and Huang 2020), denoising (Vincent et al. 2008). Besides, Pseudo label-based approaches include patch ordering (Doersch, Gupta, and Efros 2015), rotation angle prediction (Gidaris, Singh, and Komodakis 2018), frame tracking (Wang and Gupta 2015), solving the jigsaw puzzle (Noroozi and Favaro 2016).

Self-supervised Video Representation Learning

Researches in video self-supervised learning follow a similar trajectory as image self-supervised learning, which also have two groups of approaches, i.e., pretext tasks and contrastive learning. Video pretext tasks explored natural video properties or statistics as supervision signal on unlabeled data, e.g., frame prediction (Vondrick, Pirsiavash, and Torralba 2016; Behrmann, Gall, and Noroozi 2021; Luo et al. 2017), spatio-temporal puzzling (Kim, Cho, and Kweon 2019), video statistics (Wang et al. 2021b), temporal ordering (Misra, Zitnick, and Hebert 2016; Yao et al. 2021), video playback rate prediction (Benaim et al. 2020; Jenni, Meishvili, and Favaro 2020), temporal consistency (Wang, Jabri, and Efros 2019; Jabri, Owens, and Efros 2020). Recently, inspired by the success of contrastive learning on static images, contrastive learning was expanded in video

self-supervised learning (Chen et al. 2021; Alwassel et al 2019; Sermanet et al. 2018; Liu et al. 2021).

Despite the success of contrastive learning and playback rate prediction, contrastive learning approaches just focus or discriminating instances by the similarity of features but ignore the intermediate state of learned representation such as the similarity degree of features, which limits the overall performance. Besides, although pretext tasks and contrastive learning were proven to be successful in video self-supervised learning, few efforts have been placed on the mutual influence of multiple contrastive learning schemes and pretext tasks. Also, there is no guidance for a joint optimization. To settle the first problem, we proposed a novel pretext task, i.e., spatio-temporal overlap rate prediction, to encourage a model to learn the degree of similarity so as to enhance spatio-temporal feature leanring. For the second problem, based on the contrastive learning, we proposed a joint learning with rotation angle prediction to further enhance the spatial representations. Besides, we explore the mutual influence of contrastive learning and provide guidance for the joint optimization.

Methodology

Preliminary-Contrastive Learning Frameworks

In this subsection, we briefly review four relative state-of-the-art contrastive learning frameworks in this paper. Let $X = \{\mathbf{x_i}\}, i \in [1, N]$ denotes the video training dataset, where N is the total number of videos in the dataset.

SimCLR (Chen et al. 2020) SimCLR consists of four major steps:

Data Augmentation. A stochastic spatio-temporal data augmentation $\mathcal{T}$ is performed on each input $\{\mathbf{x}_i\}$, $i \in [1, N_B]$ in a minibatch, N_B is the minibatch size. Two correlated views $\mathbf{v}_i$, $\mathbf{v}'_i$ are generated from $\mathbf{x}_i$ via by $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$. The generated pair $(\mathbf{v}_i, \mathbf{v}'_i)$ is considered as a positive pair. Generated pairs $(\mathbf{v}_i, \mathbf{v}_j)$, $(\mathbf{v}_i, \mathbf{v}'_j)$ are considered as negative pairs, where $i \in [1, N_B], j \neq i$.

Feature Encoding. A spatio-temporal CNN model f_θ that extracts representation vectors from augmented data examples $\{\mathbf{v}_i, \mathbf{v'}_i\}$, $i \in [1, N_B]$, where θ is the parameter of the model.

Latent Space Projection. A multi-layer perception (MLP) projection head g_θ that maps the representation $\mathbf{h}_i$ to a latent space $\mathbf{z}_i$ which is used for contrastive loss calculation.

Contrastive Loss. With the obtained projected features z_i, $i \in [1, N_B]$, InfoNCE (Oord, Li, and Vinyals 2018) is adopted as the contrastive objective:

$$\mathcal{L}_{ctr} = -\sum_{i=1}^{N_B} \log \frac{\exp\left(sim\left(\mathbf{v}_i, \mathbf{v'}_i\right)/\alpha\right)}{\sum_{k \in \{\mathbf{v}_j, \mathbf{v'}_j\}, j \neq i} \exp\left(sim\left(\mathbf{v}_i, k\right)/\alpha\right)}, \tag{1}$$

where $sim\left(\mathbf{x}, \mathbf{y}\right) = \frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{x}||||\mathbf{y}||}$ that is the cosine similarity calculation of two input vectors.

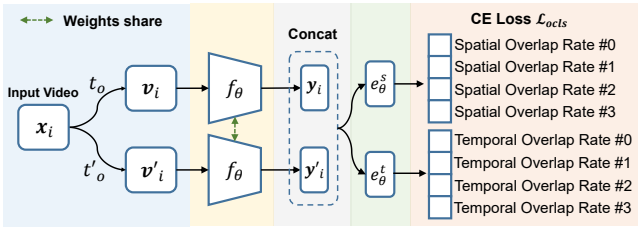


Figure 2: An overview of spatio-temporal overlap rate (STOR) prediction.

MoCo (He et al. 2020) The architecture of MoCo is similar to SimCLR. The difference between MoCo and SimCLR is that, MoCo builds a queue to store negative samples while SimCLR regards the embeddings in the mini-batch as negative samples. Besides, MoCo uses an explicit momentum encoder which adopts a moving average manner with a momentum parameter m.

$$\xi \leftarrow m\xi + (1-m)\theta, \tag{2}$$

where θ is the parameter in encoder f_θ, ξ is the parameter in momentum encoder f_ξ.

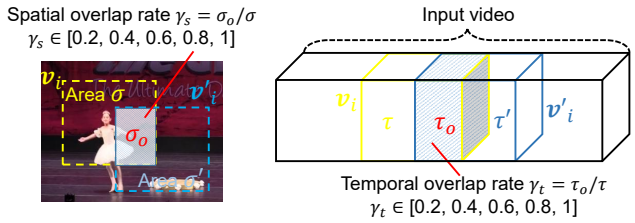
BYOL (Grill et al. 2020) BYOL is a typical contrastive learning method that does not use negative samples in the contrastive loss. BYOL can be seen as a form of MoCo without negative samples but uses an extra MLP predictor p_θ after the latent space $\mathbf{z}_i$ for predicting latent space z'_i to learn the data representation. The objective of BYOL is to minimize the distance between predicted latent features q_i from normal encoder and the projected latent feature z'_i of momentum encoder, which is:

$$\mathcal{L}_{ctr} = -\sum_{i=1}^{N_B} sim(\mathbf{v}_i, \mathbf{v'}_i), \tag{3}$$

SimSiam (Chen and He 2021) SimSiam can be seen as a form of BYOL that does not use a momentum encoder. The difference is that SimSiam uses a Siamese encoder to replace the momentum encoder, but only backpropagate one of the two encoders, i.e., stopping the gradients of the second encoder.

Spatio-temporal Overlap Rate Prediction

Since contrastive learning approaches and existing pretext tasks do not percept the degree of similarity of generated samples while humans can make it (Figure 1). Following the nature of humans, we propose a novel pretext task spatio-temporal overlap rate (STOR) prediction to help models understand the similarity of degree to further enhance the spatio-temporal representation. Following the settings of pretext tasks, our proposed approach encourages CNN model to discriminate different spatio-temporal overlap rates of two training video clips to learn the video representations. The hypothesis is that the network can only succeed in such a spatio-temporal overlap reasoning task when it understands the underlying video content and learns representative spatio-temporal features. The pipeline of STOR can be seen in Figure 2.



(a) Spatial overlap rate prediction (b) Temporal overlap rate prediction

Figure 3: Conceptual demonstration of spatio-temporal overlap rate prediction. σ denotes the spatial area of bounding box, $\sigma = \sigma'$. σ_o denotes spatial overlap area. τ denotes the temporal overlap duration, $\tau = \tau'$. τ_o denotes temporal overlap duration.

The spatio-temporal overlap rate prediction task consists of two parts, i.e., spatial overlap rate prediction and temporal overlap rate prediction. The conceptual demonstration of spatial and temporal overlap rates can be seen in Figure 3 (a) and (b). When generating samples $\mathbf{v}_i$ and $\mathbf{v'}_i$, spatio-temporal overlap data augmentation will randomly pick a spatial overlap rate $\gamma_s = \frac{\sigma_o}{\sigma}$ and a temporal overlap rate $\gamma_t = \frac{\tau_o}{\tau}$, where σ denotes the spatial area of bounding box, σ_o is the overlap area, τ denotes the temporal overlap duration and τ_o is the temporal overlap duration. Then, it generates two clips with the sampled spatial and temporal overlap rates.

Formally, we denote the STOR as $t_o \sim \mathcal{T}_o$. Given a video $\mathbf{x}_i$ and the neural network f_θ, two random transformations $t_o(\mathbf{x}_i)$ and $t'_o(\mathbf{x}_i)$ are applied to obtain the training clip $\mathbf{v}_i, \mathbf{v'}_i$, respectively. The randomly sampled spatial overlap rate is s_o^s and the randomly sampled temporal overlap rate is s_o^t. Then, we conduct feature extraction and obtain the hidden feature $\mathbf{y}_i = f_\theta(\mathbf{v}_i)$, $\mathbf{y'}_i = f_\theta(\mathbf{v'}_i)$, where the same CNN model f_θ is used. Furthermore, following the design of multiple tasks for self-supervised learning (Chen et al. 2020), two individual fully-connected (FC) heads e_θ^s and e_θ^t are adopted as the classifiers for spatial overlap rate prediction and temporal overlap rate prediction, respectively. $a_o^s = e_o^s(\mathbf{y}_i)$ and $a_o^t = e_o^t(\mathbf{y'}_i)$ are used to make a prediction of the sampling label s_o^s and s_o^t, respectively. The prediction probabilities are $p_o^s = \frac{\exp a_o^s}{\sum_o^{N_o^s} \exp a_o^s}$, $p_o^t = \frac{\exp a_o^t}{\sum_o^{N_o^t} \exp a_o^t}$, where N_o^s and N_o^t are the number of all the spatial overlap rate and temporal overlap rate candidates, respectively The parameters of the neural network f_θ are trained with a joint cross entropy (CE) loss $\mathcal{L}_{ocls}$ described as:

$$\mathcal{L}_{ocls} = -l_o^s \sum_{o=1}^{N_o^s} s_o^s \log p_o^s - l_o^t \sum_{o=1}^{N_o^t} s_o^t \log p_o^t, \qquad (4)$$

where l_o^s, l_o^t are the weights for spatial and temporal overlap tasks.

Spatio-temporal Overlap Data Augmentation

A spatio-temporal overlap data augmentation is proposed in this work for aforementioned pretext tasks. To implement the proposed spatio-temporal overlap rate prediction task, it is critical to obtain two video clips with overlap rate while keeping the random distribution of the input video clip samples. In the spatio-temporal overlap data augmentation, a 2-step sampling strategy is designed for training. Following previous video self-supervised learning method (Wang et al. 2021a; Han, Xie, and Zisserman 2020b), we adopt base data augmentation that includes multi-scale random cropping, random color jittering, random temporal jittering, random rotation jittering (randomly rotate 0-10), random playback rates and random rotation (random rotate $0°, 90°, 180°, 270°$). Given a video $\mathbf{x}_i$, we generate the first video clip sample $\mathbf{v}_i$ with the base random data augmentation in a video. Then, we randomly pick a spatial overlap rate s_o^s and a temporal overlap rate s_o^t, respectively. After that, to keep the same aspect ratio of the first video clip $\mathbf{v}_i$, the second video clip sample $\mathbf{v'}_i$ adopts the same spatial size (i.e. width w and height h) and the same playback rate of the first video clip sample $\mathbf{v}_i$. Then, the second video clip sample $\mathbf{v'}_i$ randomly crops spatial area with size (w, h) according to the picked overlap rate s_o^s and randomly clips temporal duration according to the temporal overlap rate s_o^t. After that, base data augmentation except from random cropping and temporal jittering is implemented on $\mathbf{v'}_i$. In this way, the video clip samples $\mathbf{v}_i$ and $\mathbf{v'}_i$ have a spatio-temporal overlap rate with each other, meanwhile, follow a random distribution of data augmentation.

Full Scheme of CSTP

Taking BYOL as the contrastive learning method for example, the full framework of our proposed CSTP can be seen in Figure 4. Given an input video $\mathbf{x}_i$, we first randomly generate two fixed-length clips, i.e., $\mathbf{v}_i$ and $\mathbf{v'}_i$, from different spatial and temporal locations of $\mathbf{x}_i$ according to the spatio-temporal overlap data augmentation. In this way, the two input video clips $\mathbf{v}_i$ and $\mathbf{v'}_i$ have different low-level (pixel, curve, et al.) distribution but are consistent in the semantic level. Then, the two sampled video clips $\mathbf{v}_i$ and $\mathbf{v'}_i$ are fed into a 3D CNNs f_θ and f'_ξ to extract the feature representation y_i^θ and y'^ξ_i, respectively. Besides, $\mathbf{v'}_i$ is fed into f_θ to extract feature representation y'^θ_i. Afterwards, the extracted feature y_i^θ is fed into playback rate prediction head c_θ for playback rate prediction and obtain the loss $\mathcal{L}_{pcls}$; The extracted feature y_i^θ is fed into rotation prediction head d_θ for rotation prediction and obtain loss the loss $\mathcal{L}_{rcls}$; The extracted features y_i^θ and y'^θ_i are concatenated and the concatenated features $cat(y_i^\theta, y'^\theta_i)$ are fed into e_θ^s and e_θ^t for spatial overlap rate prediction and temporal overlap rate prediction, respectively, and obtain the loss $\mathcal{L}_{ocls}$ as Equation (4); The three losses $\mathcal{L}_{pcls}$, $\mathcal{L}_{rcls}$, $\mathcal{L}_{ocls}$ are all CE loss. The extracted features y_i^θ and y'^θ_i follow the steps of contrastive learning scheme and obtain the loss $\mathcal{L}_{ctr}$. Lastly, contrastive loss and three spatio-temporal pretext loss are jointed for optimization. The final loss is as:

$$\mathcal{L} = \lambda_{ctr}\mathcal{L}_{ctr} + \lambda_{pcls}\mathcal{L}_{pcls} + \lambda_{rcls}\mathcal{L}_{rcls} + \lambda_{ocls}\mathcal{L}_{ocls}, \qquad (5)$$

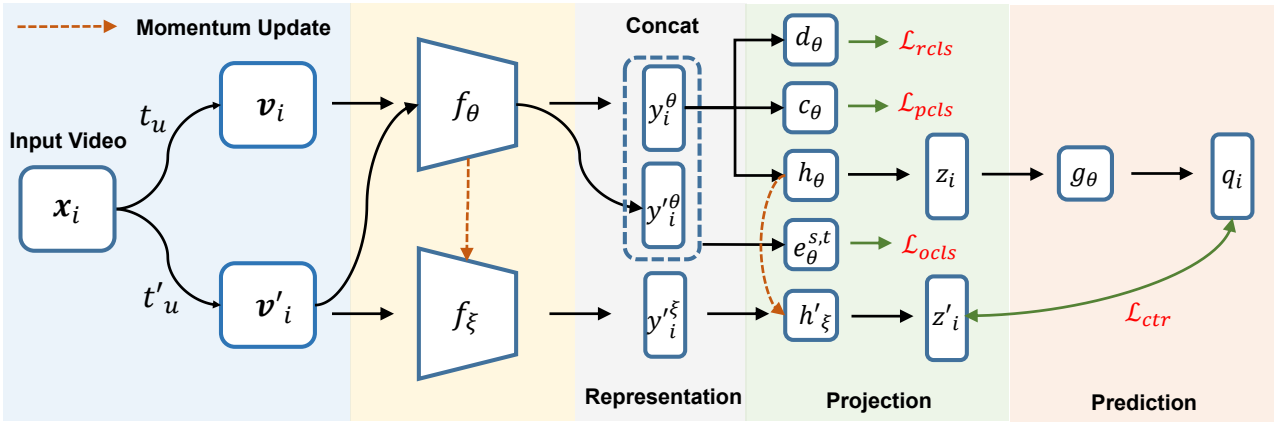


Figure 4: An overview of contrastive spatio-temporal pretext (CSTP) approach for video self-supervised learning. The approach consists of contrastive learning and three pretext tasks (playback rate prediction, rotation prediction, spatio-temporal overlap rate (STOR) prediction).

where λ_{ctr}, λ_{pcls}, λ_{rcls}, λ_{ocls} are hyperparameters that control the weights of the regularization term.

Experiments

Datasets

Kinetics-400 (Carreira and Zisserman 2017) is one of the large-scale action recognition benchmarks which contains around 300K videos over 400 action categories. **UCF-101** (Soomro, Zamir, and Shah 2012) is a widely used benchmark for action recognition. It has three splits, which consists of 13, 320 videos that cover 101 human action classes. **HMDB-51** (Kuehne et al. 2011) is also a small-scale dataset for action recognition, which consists of three splits and 6, 770 videos in 51 actions.

Ablations Studies

Understanding of transformation-based pretext tasks. 3D Rot (Jing et al. 2018), RTT (Jenni, Meishvili, and Favaro 2020) has verified that video transformation-based pretext task can benefit video self-supervised learning. It was claimed that the gain on performance comes from the pseudo labels learning, however, we found that the gain in performance of transformation-based pretext task comes from two aspects, i.e., more diverse data (data augmentation along with pretext task) and the pseudo label prediction.

In this study, we adopt R(2+1)D as the backbone and conduct experiments on split 1 of dataset UCF-101. The results are summarized in Table 1. In the table, "Base" means basic data augmentation methods which includes multi-scale random cropping, random gaussian blur, random color jittering, random temporal jittering. Exp. 1 denotes the baseline performance, which was trained from scratch (randomly initializing weights) with the aforementioned data augmentation methods. We can observe that when the model was trained from scratch with basic data augmentation methods, the baseline performance is 60.3% (Exp. 1). After adding playback rate data augmentations, the performance increased to 67.5% (+7.2%, Exp. 2). After further combining playback rate prediction pretext task, the performance increased to 77.2 (+9.7%, Exp. 5). The same trend can be seen in the rotation data augmentation and pretext task (Exp. 3&7 vs Exp. 1). Besides, we observe that combining the playback rate data augmentation and the rotation data augmentation can benefit the performance (Exp. 2&3 vs Exp. 4). In terms of our proposed STOR prediction task, it should be noted that unlike playback rates and rotation operation, our proposed overlap operations will not increase the diversity of data. When performing STOR with base data augmentation, the performance increased to 69.6% (+9.3%). The gain in performance is comparable to playback rate prediction and rotation prediction, which verifies the effectiveness of our proposed STOR. And we also can observe the further gain with additional data augmentation methods (Exp. 10&11 vs Exp. 9).

Exploration of CSTP Learning Methods. To verify the effectiveness of our proposed CSTP and figure out the impact of each contribution, experiments on each part of CSTP were conducted. We take BYOL as the contrastive learning framework for exploration. R(2+1)D was adopted as the backbone in the experiments. The performance of each part in CSTP is shown in Table 2. Firstly, we can observe that each part of CSTP can boost the performance of the baseline. The baseline performance is 71.1% (scratch). The performance was further increased to 76.2% (+5.1%, Exp. 2), 78.7% (+7.6%, Exp. 3), 78.5% (+7.4%, Exp. 4), 76.6% (+5.5%, Exp. 5) when the training was conducted from the pre-trained model of contrastive learning, playback rate prediction, rotation prediction and overlap rate prediction, respectively. Secondly, it can be observed that combining each pretext task to contrastive learning can boost the performance of individually using contrastive learning or the pretext task. Thirdly, STOR can boost the performance of each component in CSTP. When STOR is combined with playback rate prediction and rotation prediction, the performance is improved by +3.3% and +3.1%, respectively. When com-

Exp.	Data Augmentation			Pretext Tasks			Dataset
	Base	PB Rate	Rot.	PB Rate	Rot.	STOR	UCF-101 (%)
1	✓	✗	✗	✗	✗	✗	60.3
2	✓	✓	✗	✗	✗	✗	67.5
3	✓	✗	✓	✗	✗	✗	66.9
4	✓	✓	✓	✗	✗	✗	71.1
5	✓	✓	✗	✓	✗	✗	77.2
6	✓	✓	✓	✓	✗	✗	78.7
7	✓	✗	✓	✗	✓	✗	77.6
8	✓	✓	✓	✗	✓	✗	78.5
9	✓	✗	✗	✗	✗	✓	69.6
10	✓	✓	✗	✗	✗	✓	74.1
11	✓	✓	✓	✗	✗	✓	76.2

Table 1: Exploration of transformation-based pretext tasks. Results are evaluated on split 1 of UCF-101. The backbone is R(2+1)D. "Base" means basic data augmentation methods which includes multi-scale random cropping, random gaussian blur, random color jittering, random temporal jittering.

Exp.	Contr.	Pretext Tasks			Dataset (%)	
		PB Rate	Rot.	STOR	UCF-101	HMDB-51
1	✗	✗	✗	✗	71.1	38.3
2	✓	✗	✗	✗	76.6	45.5
3	✗	✓	✗	✗	78.7	48.2
4	✗	✗	✓	✗	78.5	47.1
5	✗	✗	✗	✓	76.2	45.4
6	✗	✓	✓	✗	84.1	53.5
7	✗	✓	✗	✓	82.0	51.7
8	✗	✗	✓	✓	81.6	51.4
9	✓	✓	✗	✗	81.8	51.4
10	✓	✗	✓	✗	81.5	51.1
11	✓	✗	✗	✓	77.7	49.5
12	✗	✓	✓	✓	85.3	54.9
13	✓	✓	✓	✗	84.4	54.1
14	✓	✓	✗	✓	83.1	53.8
15	✓	✗	✓	✓	82.6	52.7
16	✓	✓	✓	✓	**85.6**	**55.1**

Table 2: Exploration of each part of contrastive spatio-temporal pretext learning methods. Results are evaluated on split 1 of UCF-101 and HMDB-51.

bining STOR with contrastive learning, the performance is improved by 1.1%. The reason why the improvement on contrastive learning is less than pretext task may be that S-TOR and contrastive learning both are learning the similarity of samples, which is independent to playback rate prediction and rotation prediction. Finally, through comparisons (Exp. 6 vs Exp. 3&4; Exp. 7 vs Exp. 3&5, etc.), it can be observed that in CSTP, the combination of multiple pretext tasks and contrastive learning can improve the overall performance, which verifies the importance of the learned intermediate state in representation.

Exploration of Contrastive Learning Methods. To explore the mutual influence of multiple contrastive learning schemes and different pretext tasks, we conducted experiments on four popular contrastive learning schemes which are SimCLR, MoCo, BYOL and SimSiam. In Table 3, we compared four contrastive learning schemes in the context of CSTP. Except from the contrastive learning scheme, the settings of the experiments in Table 3 are the same as those in Table 2. Firstly, we observe that four contrastive learning schemes can work with the pretext tasks in CSTP by comparing the results of each contrastive learning schemes in Table 3 with the results of only pretext tasks in Table 2. For example, when only using SimCLR in CSTP, the performance is 75.5%, while the performance increased to 80.7% (+5.2%), 82.9% (+7.4%) and 84.4% (+8.9%) as gradually combining playback rate prediction, rotation prediction and STOR prediction. Secondly, in the four contrastive learning schemes, we observe that BYOL performed the best in our proposed CSTP framework. Thirdly, we observe the four contrastive learning schemes share a similar trend when combining different pretext tasks. Besides, we observe that the four contrastive learning schemes share a similar trend when combining with pretext tasks. Finally, we can observe that momentum update is useful for CSTP (comparing performance of BYOL and SiaSiam, BYOL outperforms SiaSiam); negative samples are not critical for CSTP when using momentum update (comparing performance of BYOL and MoCo, BYOL outperforms MoCo).

Exploration of STOR. From Table 2&3, we observe that combining STOR benefits the overall results in CSTP. STOR consists of two pretext tasks, e.g., spatial overlap rate prediction and temporal overlap prediction. To demonstrate the effectiveness of these two pretext tasks, we separate STOR apart and conduct experiments on each part. To remove the influence of the data augmentation methods of other pretext tasks, "Base" data augmentation is only used in the experiments. The results are summarized in Table 5. When training from scratch, the result in UCF-101 is 60.3%. While after employing spatial overlap prediction pretext task, the performance increased to 67.8% (+7.5%). Similar trend can be observed in temporal overlap prediction, which increased the baseline to 66.5% (+6.2%). Besides, combining the spatial overlap prediction and temporal overlap prediction can further increase the performance to 69.6% (+9.3%).

Exploration of different spatio-temporal rates in S-TOR. In this experiment, we explore the influence of different candidates to STOR pre-training strategy. R(2+1)D was adopted as the backbone. We only use STOR as the pre-training strategy and fine-tune the whole model as demonstrated in Experimental Settings in Appendix. We conducted 6 sets of candidates to demonstrate the influence of the choice of candidates, which are 2 candidates

Contr.	Pretext Tasks			Dataset (%)	
	PB Rate	Rot.	STOR	UCF-101	HMDB-51
SimCLR	✗	✗	✗	75.5	44.3
	✓	✗	✗	80.7	50.1
	✓	✓	✗	82.9	52.8
	✓	✓	✓	84.4	54.0
MoCo	✗	✗	✗	75.8	44.8
	✓	✗	✗	81.3	52.8
	✓	✓	✗	83.5	53.4
	✓	✓	✓	84.9	54.3
BYOL	✗	✗	✗	76.6	45.5
	✓	✗	✗	81.8	51.4
	✓	✓	✗	84.4	54.1
	✓	✓	✓	**85.6**	**55.1**
SimSiam	✗	✗	✗	75.2	44.2
	✓	✗	✗	80.6	50.3
	✓	✓	✗	82.8	52.6
	✓	✓	✓	83.9	53.6

Table 3: Exploration of contrastive learning methods of CSTP. Results are evaluated on split 1 of UCF-101 and HMDB-51.

Number of candidates	UCF-101 (%)	HMDB-51 (%)
2	72.5	42.5
3	72.4	43.3
4	73.9	44.8
5	**76.2**	**45.4**
6	75.4	45.0
7	75.5	44.9

Table 4: Exploration of STOR. Results are evaluated on split 1 of UCF-101. Results without data augmentation methods of other pretext tasks.

Spatial Overlap rate Prediction	Temporal Overlap rate Prediction	UCF-101 (%)
✗	✗	60.3 (baseline)
✓	✗	67.8
✗	✓	66.5
✓	✓	**69.6**

Table 5: Exploration of STOR. Results are evaluated on split 1 of UCF-101. Results without data augmentation methods of other pretext tasks.

- $[0.5, 1]$, 3 candidates - $[0.33, 0.66, 0.99]$, 4 candidates - $[0.25, 0.5, 0.75, 1.0]$, 5 candidates - $[0.2, 0.4, 0.6, 0.8, 1.0]$, 6 candidates - $[0.166, 0.332, 0.498, 0.664, 0.83, 1.0]$, 7 candidates - $[0.143, 0.286, 0.429, 0.572, 0.715, 0.858, 1]$. The comparative results of different candidates of STOR prediction task in UCF-101 and HMDB-51 datasets are shown in Table 4.

We can observe that the two candidates in STOR performed the worst results. And with the number of candidates increased, the results tend to improve till the number of candidates reaches five. When the number of candidates is bigger than five, the performance tends to be saturated. The reason may be that few candidates make perceiving STOR prediction easy, so that the model cannot learn spatial-temporal representation well. When increasing the number of candidates, the task of STOR prediction is getting harder, which requires the model stronger representation ability. And when the number of candidates was over five, the STOR task may be too hard to obtain better representation ability.

Evaluation on Action Recognition Task

End-to-end fine-tuning stage. Following previous work (Chen et al. 2021; Jenni, Meishvili, and Favaro 2020), we compare our proposed method with recent state-of-the-art self-supervised video representation learning approaches. In Table 6, for fair comparison, we report the Top-1 accuracy of UCF-101 and HMDB-51 datasets with commonly used network backbones (C3D, R(2+1)D, S3D) and pre-training datasets, where the results are averaged over three splits of the two datasets. When trained from pre-trained model on UCF-101, it can be observed that our proposed method can achieve state-of-the-art performance among all the approaches with all the backbones. Specifically, our C-STP achieves 85.7%/55.3% with R(2+1)D backbone, which outperforms state-of-the-art performance by 4.1%/6.1% in UCF-101/HMDB-51 datasets. It even outperforms some approaches trained on Kinetic-400. Similar trend can be observed on C3D and S3D. This demonstrates that our proposed method is generalized to multiple network architectures, which has a robust spatio-temporal modeling ability. When trained from pre-trained model on Kinetics-400, we observe that our CSTP could achieve the state-of-the-art performance in on all the backbones with same input modality. Notably, CSTP unsupervised pre-training leads the accuracy by 0.6%/1.6% against fully-supervised ImageNet pre-training with only pre-trained on the small dataset UCF-101.

Linear Evaluation. We also follow the settings of previous work (Jenni, Meishvili, and Favaro 2020) for the linear evaluation in which we freeze the convolutional layers and only train the final FC layers for classification. The Freeze column in Table 6 denotes the linear evaluation setting. We can observe that our method obtains state-of-the-arts performance on UCF-101 and HMDB-51. Specifically, our approach also outperforms TaCo (Bai et al. 2020) which combined four existing pretexts tasks with MoCo, demonstrating the representation ability of our proposed method.

Method	Backbone	Dataset	Freeze	Datasets (%)	
				UCF	HMDB
CCL (Kong et al. 2020)	R3D-18	K400	✓	54.0	29.5
MemDPC (Han, Xie, and Zisserman 2020a)	R3D-34	K400	✓	54.1	30.5
TaCo (Bai et al. 2020)	R3D-18	K400	✓	59.6	26.7
MFO (Qian et al. 2021)	R3D-18	K400	✓	63.2	33.4
Ours	R3D-18	K400	✓	**70.5**	**34.4**
Random Initialization (scratch)	R(2+1)D	-	✗	60.3	27.6
Supervised ImageNet Pre-training	R(2+1)D	ImageNet	✗	85.1	53.7
VCP (Luo et al. 2020)	R(2+1)D	UCF-101	✗	66.3	32.2
PRP (Yao et al. 2020)	R(2+1)D	UCF-101	✗	72.1	35.0
VCOP (Xu et al. 2019)	R(2+1)D	UCF-101	✗	72.4	30.9
PacePred (Wang, Jiao, and Liu 2020)	R(2+1)D	UCF-101	✗	75.9	35.9
RSPNet (Chen et al. 2021)	R(2+1)D	K400	✗	81.1	44.6
RTT (Jenni, Meishvili, and Favaro 2020)	R(2+1)D	UCF-101	✗	81.6	46.4
VideoMoCo (Pan et al. 2021)	R(2+1)D	K400	✗	78.7	49.2
Ours	R(2+1)D	UCF-101	✗	85.7	55.3
Ours	R(2+1)D	K400	✗	**87.6**	**56.4**
RTT (Jenni, Meishvili, and Favaro 2020)	C3D	UCF-101	✗	68.3	38.4
PRP (Yao et al. 2020)	C3D	UCF-101	✗	69.1	34.5
VCP (Luo et al. 2020)	C3D	UCF-101	✗	68.5	32.5
VCOP (Xu et al. 2019)	C3D	UCF-101	✗	65.6	28.4
MoCo+BE (Wang et al. 2021a)	C3D	UCF-101	✗	72.4	42.3
RSPNet (Chen et al. 2021)	C3D	K400	✗	76.7	44.6
Ours	C3D	UCF-101	✗	80.8	48.9
Ours	C3D	K400	✗	**82.6**	**50.2**
MFO (Qian et al. 2021)	S3D	UCF-101	✗	74.3	37.2
CoCLR* (Han, Xie, and Zisserman 2020b)	S3D	UCF-101	✗	81.4	52.1
CoCLR* (Han, Xie, and Zisserman 2020b)	S3D	K400	✗	**87.9**	54.6
Ours	S3D	UCF-101	✗	83.6	53.3
Ours	S3D	K400	✗	87.3	**55.7**

*CoCLR used optical flow as guidance on training.

Table 6: Comparison With State-of-the-arts. Average results of three splits in UCF-101 and HMDB-51 datasets.

Evaluation on Video Retrieval Task

Besides the video action recognition task, we also report the Top-k ($k = 1, 5, 10, 20, 50$) video retrieval performance with two different backbones, e.g., R(2+1)D and C3D. The quantitative results are shown in Table 7. We observe that our proposed approach outperforms state-of-the-art approaches by a large margin under different k with the two network backbones. Specifically, our CSTP outperforms the state-of-the-art model by 16.6%/6.9% in UCF-101/HMDB-51 dataset using R(2+1)D as the backbone, and outperforms the state-of-the-art model by 14.7%/6.7% using C3D as the backbone. These imply that our proposed CSTP could help the model to learn more discriminating spatio-temporal features.

Qualitative analysis

We further provide retrieval results as a qualitative study for the proposed CSTP as shown in Figure 5. We can see that for the listed five query clips, videos with similar motion characteristics and appearance are successfully retrieved. This implies that CSTP can learn both meaningful motion and appearance representations for videos. We also use the class

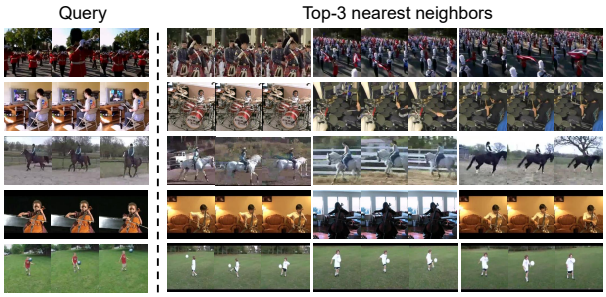


Figure 5: Qualitative examples of video retrieval. Retrieval of Top-3 similar samples. The videos are "BandMarching", "Drumming", "HorseRiding", "PlayingCello", "SoccerJuggling", respectively from top to bottom.

Method	Backbone	UCF-101 (%)			HMDB-51 (%)		
		R@1	R@5	R@10	R@1	R@5	R@10
RSPNet [2021]	R3D-18	41.1	59.4	68.4	-	-	-
MFO [2021]	R3D-18	39.6	57.6	69.2	18.8	39.2	51.0
VCOP [2019]	R(2+1)D	10.7	25.9	35.4	5.7	19.5	30.7
VCP [2020]	R(2+1)D	19.9	33.7	42.0	6.7	21.3	32.7
PRP [2020]	R(2+1)D	20.3	34.0	41.9	8.2	25.3	36.2
PacePred [2020]	R(2+1)D	25.6	42.7	51.3	12.9	31.6	43.2
IIC [2020]	R(2+1)D	34.7	51.7	60.9	12.7	33.3	45.8
Ours	R(2+1)D	**51.3**	**71.1**	**80.3**	**21.6**	**48.4**	**62.4**
VCOP [2019]	C3D	12.5	29.0	39.0	7.4	22.6	34.4
VCP [2020]	C3D	17.3	31.5	42.0	7.8	23.8	35.3
PRP [2020]	C3D	23.2	38.1	46.0	10.5	27.2	40.4
PacePred [2020]	C3D	31.9	49.7	59.2	12.5	32.2	45.4
IIC [2020]	C3D	31.9	48.2	57.3	11.5	31.3	43.9
RSPNet [2021]	C3D	36.0	56.7	66.5	-	-	-
MoCo+BE [2021]	C3D	-	-	-	10.2	27.6	40.5
Ours	C3D	**50.7**	**69.4**	**77.9**	**19.3**	**46.6**	**61.0**

Table 7: Recall-at-Topk. Comparison with state-of-the-art methods in video retrieval task on UCF-101 and HMDB-51.

activation map technique (CAM) (Zhou et al. 2016) to visualize the Region of Interest (RoI) following the previous works (Qian et al. 2021; Kuang et al. 2021). We visualize the activations of the output of the last convolutional layers from R(2+1)D models pretrained by contrastive learning (BYOL) and the proposed CSTP (BYOL based). As seen in Figure 6, CSTP can focus more on discriminative motion regions, while contrastive learning cannot perceive important motion cues well. For example, our approach precisely focuses on the moving hands in the playing guitar scene, while contrastive learning approach regards the body and some background as the RoI.

Figure 6: Visualization of RoI learn for contrastive learning and CSTP. Our model focuses more on the regions containing motion and appearance information for two pretext tasks. The videos are "ApplyEyeMakeUp", "HorseRiding", "PlayingGuitar", "BrushingTeeth", respectively from left to right

Conclusion

Contrastive learning focuses on discriminating the similarity of features while ignoring the intermediate state of learned representations. In this paper, considering the degree of similarity as the intermediate state, we propose a new pretext task - spatio-temporal overlap rate prediction (STOR) in a way of inter-feature reference, which enhances spatio-temporal representation learning by discriminating the STOR of two relative samples. Besides, we propose a joint optimization framework contrastive spatio-temporal pretext (CSTP) to further enhance spatio-temporal feature learning. Furthermore, we study the mutual influence of each component in CSTP and provide design guidance. Extensive experiments show that the STOR prediction task can benefit self-supervised learning. Besides, CSTP framework is flexible and achieves state-of-the-art performance on action recognition and video retrieval downstream tasks with different backbones. It is worth noting that the way of intermediate state perception and inter-feature reference in STOR provides new perspectives for self-supervised learning community.

References

Alwassel, H.; Mahajan, D.; Korbar, B.; Torresani, L.; Ghanem, B.; and Tran, D. 2019. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*.

Bai, Y.; Fan, H.; Misra, I.; Venkatesh, G.; Lu, Y.; Zhou, Y.; Yu, Q.; Chandra, V.; and Yuille, A. 2020. Can Temporal Information Help with Contrastive Self-Supervised Learning? *arXiv preprint arXiv:2011.13046*.

Behrmann, N.; Gall, J.; and Noroozi, M. 2021. Unsupervised Video Representation Learning by Bidirectional Feature Prediction. In *Proc. WACV*, 1670–1679.

Benaim, S.; Ephrat, A.; Lang, O.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; Irani, M.; and Dekel, T. 2020. Speednet: Learning the speediness in videos. In *Proc. CVPR*, 9922–9931.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. CVPR*, 6299–6308.

Chen, P.; Huang, D.; He, D.; Long, X.; Zeng, R.; Wen, S.; Tan, M.; and Gan, C. 2021. Rspnet: Relative speed perception for unsupervised video representation learning. In *Proc. AAAI*, volume 1.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proc. CVPR*, 15750–15758.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proc. ICCV*, 1422–1430.

Dosovitskiy, A.; Fischer, P.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2015. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(9): 1734–1747.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proc. ICCV*, 6202–6211.

Feichtenhofer, C.; Fan, H.; Xiong, B.; Girshick, R.; and 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*, 3299–3309.

Fernando, B.; Bilen, H.; Gavves, E.; and Gould, S. 2017. Self-supervised video representation learning with odd-one-out networks. In *Proc. CVPR*, 3636–3645.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Han, T.; Xie, W.; and Zisserman, A. 2020a. Memory-augmented dense predictive coding for video representation learning. In *Proc. ECCV*, 312–329. Springer.

Han, T.; Xie, W.; and Zisserman, A. 2020b. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, 9729–9738.

Jabri, A.; Owens, A.; and Efros, A. A. 2020. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*.

Jenni, S.; Meishvili, G.; and Favaro, P. 2020. Video representation learning by recognizing temporal transformations. In *Proc. ECCV*, 425–442. Springer.

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

Kim, D.; Cho, D.; and Kweon, I. S. 2019. Self-supervised video representation learning with space-time cubic puzzles. In *Proc. AAAI*, volume 33, 8545–8552.

Kong, Q.; Wei, W.; Deng, Z.; Yoshinaga, T.; and Murakami, T. 2020. Cycle-contrast for self-supervised video representation learning. *arXiv preprint arXiv:2010.14810*.

Kuang, H.; Zhu, Y.; Zhang, Z.; Li, X.; Tighe, J.; Schwertfeger, S.; Stachniss, C.; and Li, M. 2021. Video Contrastive Learning with Global Context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3195–3204.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Liu, Y.; Wang, K.; Lan, H.; and Lin, L. 2021. Temporal Contrastive Graph for Self-supervised Video Representation Learning. *arXiv e-prints*, arXiv–2101.

Luo, D.; Liu, C.; Zhou, Y.; Yang, D.; Ma, C.; Ye, Q.; and Wang, W. 2020. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11701–11708.

Luo, Z.; Peng, B.; Huang, D.-A.; Alahi, A.; and Fei-Fei, L. 2017. Unsupervised learning of long-term motion dynamics for videos. In *Proc. CVPR*, 2203–2212.

Misra, I.; Zitnick, C. L.; and Hebert, M. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *Proc. ECCV*, 527–544. Springer.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. ECCV*, 69–84. Springer.

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

Pan, T.; Song, Y.; Yang, T.; Jiang, W.; and Liu, W. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proc. CVPR*, 11205–11214.

Qian, R.; Li, Y.; Liu, H.; See, J.; Ding, S.; Liu, X.; Li, D.; and Lin, W. 2021. Enhancing Self-supervised Video Representation Learning via Multi-level Feature Optimization. *arXiv preprint arXiv:2108.02183*.

Sermanet, P.; Lynch, C.; Chebotar, Y.; Hsu, J.; Jang, E.; Schaal, S.; Levine, S.; and Brain, G. 2018. Time-contrastive networks: Self-supervised learning from video. 1134–1141. IEEE.

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

Su, J.-W.; Chu, H.-K.; and Huang, J.-B. 2020. Instance-aware image colorization. In *Proc. CVPR*, 7968–7977.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proc. ICCV*, 4489–4497.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proc. CVPR*, 6450–6459.