Video Playback Rate Perception for Self-supervised Spatio-Temporal Representation Learning

Yuan Yao[1]*, Chang Liu[1]*, Dezhao Luo[2], Yu Zhou[2] and Qixiang Ye[1]†

[1]University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{yaoyuan17,liuchang615}@mails.ucas.ac.cn, {luodezhao,zhouyu}@iie.ac.cn
qxye@ucas.ac.cn

Abstract

In self-supervised spatio-temporal representation learning, the temporal resolution and long-short term characteristics are not yet fully explored, which limits representation capabilities of learned models. In this paper, we propose a novel self-supervised method, referred to as video Playback Rate Perception (PRP), to learn spatio-temporal representation in a simple-yet-effective way. PRP roots in a dilated sampling strategy, which produces self-supervision signals about video playback rates for representation model learning. PRP is implemented with a feature encoder, a classification module, and a reconstructing decoder, to achieve spatio-temporal semantic retention in a collaborative discrimination-generation manner. The discriminative perception model follows a feature encoder to prefer perceiving low temporal resolution and long-term representation by classifying fast-forward rates. The generative perception model acts as a feature decoder to focus on comprehending high temporal resolution and short-term representation by introducing a motion-attention mechanism. PRP is applied on typical video target tasks including action recognition and video retrieval. Experiments show that PRP outperforms state-of-the-art self-supervised models with significant margins. Code is available at github.com/yuanyao366/PRP.

1. Introduction

Deep networks, *i.e.*, Convolutional Neural Networks (CNNs) [22], have achieved unprecedented success in computer vision area. This can be largely attributed to the learned rich representation incorporating both low-level fine-details and high-level semantics [35]. To realize rich representation, networks are typically pre-trained using large-scale image/video datasets (*e.g.*, ImageNet [16] and Kinetics [18]) under accurate annotation supervision [19].

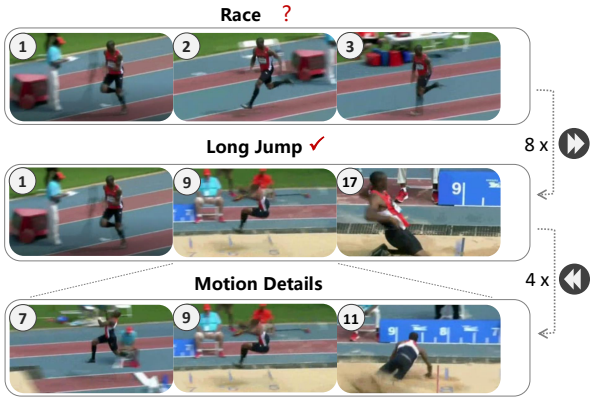

Figure 1. With limited visible frames, video clips with different playback rates (temporal resolutions) imply different semantics. A video clip with normal playback rate (first row) can be misunderstood as "race". With higher playback rate (second row), we can see that it is in fact "long jump", of which short-term motion details can be perceived in the slow-down video (third row). Perceiving videos with different playback rates is crucial in learning long-short term spatio-temporal representation.

However, large-scale data annotation is laborious, expensive, or can be impractical, particularly for complex data such as videos and concepts such as action analysis and video retrieval [10, 18]. Considering the availability of large-scale unlabelled data on the Web, self-supervised representation learning, which leverages intrinsic correspondence within unlabelled data to pre-train desired representation models, has attracted increasing attention.

Self-supervised representation learning defines an annotation-free proxy task, which leverages easily developed supervision signals from data itself to train network

*Equal contribution
†Corresponding author

models, which then facilitate the implementation of the downstream target tasks. From the perspective of frame content perception, early self-supervised methods focused on predicting the spatial transformation of images [10]. Without considering the temporal relations, however, the learned features are merely on a frame-by-frame basis, which are inappropriate to video analysis tasks because the temporal dimension defines essential differences between a video sequence and an image set. Recent works [36] learned spatio-temporal representation by regressing both motion and appearance statistics. Nevertheless, without the capability to perceive temporal resolution characteristics, such a mechanism is unable to learn long-short term representation necessary for precise video understanding, Fig. 1.

In this paper, we propose a novel self-supervised approach, referred to as video Playback Rate Perception (PRP), targeting at learning representation about multiple temporal resolutions in a simple-yet-effective manner. PRP is motivated by the motion perception mechanism observed in primate visual systems [25, 26], $i.e.$, different visual cells respond to different temporal changes. M-cells are sensitive to quick and short-term changes while P-Cells focus on slower and longer-term variation. This mechanism has been explored by SlowFast networks [7] for video recognition, while we update it to a self-supervised manner to perceive multiple temporal resolutions.

To perceive temporal resolution characteristics within video data, a dilated sampling strategy is designed to produce videos with various playback rates. The original videos simulate high playback rates relative to frame-sampled videos, and content similarity between videos of different playback rates are used as a supervision signal for representation learning.

With a discriminative model, PRP can be trained to classify videos of different playback rates. With a generative model, PRP is driven to reconstruct low playback rate videos from high playback rate ones. The discriminative perception model follows a feature encoder to focus on perceiving low temporal-resolution and long-term representation by classifying fast-forward rates. The generative perception model acts as a feature decoder to focus on comprehending high temporal-resolution and short-term representation by introducing a motion-attention mechanism. Collaborative discriminative-generative perception further aggregates long-short term representation capacity, Fig. 2.

The contributions of this work include:

- A novel video Playback Rate Perception (PRP) approach is proposed to capture temporal resolution characteristics within video domain in a self-supervised manner.
- PRP is implemented with discriminative and generative perception models, which cooperatively retain spatio-temporal semantics in representation models. Furthermore, we introduce a motion attention mechanism, which drives representation to focus on meaningful foreground regions.
- We apply PRP to three kinds of 3D CNNs and two target tasks including action recognition and video retrieval, and improve the state-of-the-arts with significant margins.

2. Related Work

Self-supervised learning leverages information from unlabelled data to train models. Existing approaches usually define an annotation-free proxy task which demands a network predicting information hidden within unannotated videos. The learned models can then be applied to target tasks (eitehr supervised or unsupervised) after fine-tuning. Conventional self-supervised methods include discriminative proxy tasks such as classifying transformed images [12, 20, 6] or video content [43], and generative proxy tasks which include image inpainting [29] and video reconstruction [34, 43].

2.1. Proxy Tasks

From a broader view, proxy tasks can be constructed on top of multiple sensory data such as ego-motion [5], sound [4], and cross-modal data [17, 30, 11]. Although in this paper, we mainly review proxy tasks based on visual signals.

Spatial Representation Learning. Spatial transforms applied to images can produce supervision signals for representation learning [23]. As a representative method, the rotation-based self-supervised approach [12, 9] learns CNN features by rotating images and using rotated angles as supervision. The completion-based approach [20, 6, 13] learns image representations by predicting damaged Jigsaw puzzles. While context impainting [2] trains the CNN model to predict content of a withheld image region conditioned according to its surroundings, the image-patch matching approach [38, 42] trains a representation model to capture spatial in-variance.

Spatio-temporal Representation Learning. The large amount of video clips with rich spatio-temporal information provide various supervision signals. In [37], the temporal continuity of video frames could be used as a supervisory signal. In [27, 24], predicting orders of frames or video clips drives learning spatio-temporal representation. In [10], an odd-one-out network was proposed to identify the unrelated or odd clips from a set of otherwise related clips. To find the odd clip, the models have to learn spatio-temporal features which can discriminate similar clips. In [3], unsupervised motion segmentation

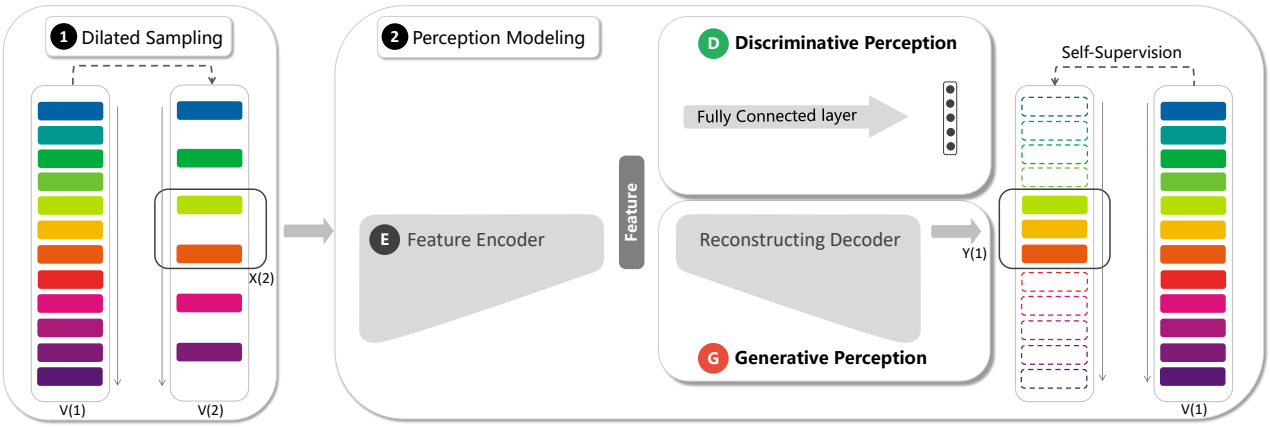

Figure 2. Playback rate perception (PRP) is composed of dilated sampling and perception modeling. Perception modeling is implemented with a feature encoder, a discriminative module, and a reconstructing decoder (generative module). The self-supervision signals are generated using dilated sampling.

on videos was used to obtain segments, which perform as pseudo ground truth to train CNNs for segmentation.

Early methods usually learn features based upon 2D CNNs and simplistically based on a frame-by-frame process, which are inappropriate to video analytic tasks where spatio-temporal features are prevailing. Recently, 3D representations are learned [36] by regressing motion and appearance statistics. The order of video clips is then used as a supervised signal for temporal representation learning [39]. 3D CNN models are trained by completing space-time cubic puzzles [19].

Despite of substantial progress in the field, existing methods unfortunately ignore the multiple temporal resolutions, which are essential for video-based tasks. Without these temporal resolution characteristics, the representation capability of learned models remains limited.

2.2. Target Tasks

For video-related tasks 3D CNN models were trained using a large-scale video databases with video category annotation [8, 32]. Nevertheless, the representation models trained on video classification tasks lack general applicability. Fine-tuning such models to other target tasks, $e.g.$, action recognition and video retrieval, could produce sub-optimal results. To conquer these issues, we propose the self-supervised PRP approach, and target at improving the model generality, by incorporating long-short term temporal representations,

3. Playback Rate Perception

Fast-forward and slow-down playback are two commonly used modes when browsing videos. To quickly understand video content, $e.g.$, a movie, we can use the fast-forward mode. To capture the fine details within a wonderful clip, we usually require action replay with a slow-down play rate. The way humans perceive video content demonstrates an important fact that the temporal resolution and long-short term characteristics are critical to get better understanding of videos.

Based on this observation, we propose the video Playback Rate Perception (PRP) for representation learning, which is composed of two components: dilated sampling and perception modeling. Dilated sampling augments video clips into different temporal resolution (fast-forward) while perception modeling learns rich spatio-temporal representation to classify videos into playback rates and/or reconstruct from the low temporal resolution videos to high temporal resolution ones (slow-down), Fig. 2.

3.1. Dilated Sampling

Given a raw video $V(1)$, we uniformly sample a video frame from each s frames with the same temporal interval, which is denoted as $s\times$ dilated sampling. This procedure generates video $V(s)$ with $s\times$ fast-forward playback rate. Considering the spatial similarity and temporal ambiguity among video frames, we sample successive l frames from $V(s)$ as a learning sample, $X(s)$, which can be fed to 3D CNNs. For the example shown in Fig. 2(left), $s = 2$ and $l = 2$. The videos $V(s)$ with different dilated sampling intervals have consistent content but different playback rates. Such playback rates, together with their corresponding video content, provide self-supervision signals for representation model learning.

3.2. Perception Modeling

Feature Encoder. To extract both spatial and temporal features, we choose C3D [32], R3D and R(2+1)D [33] as feature encoders. C3D is a natural extension from 2D

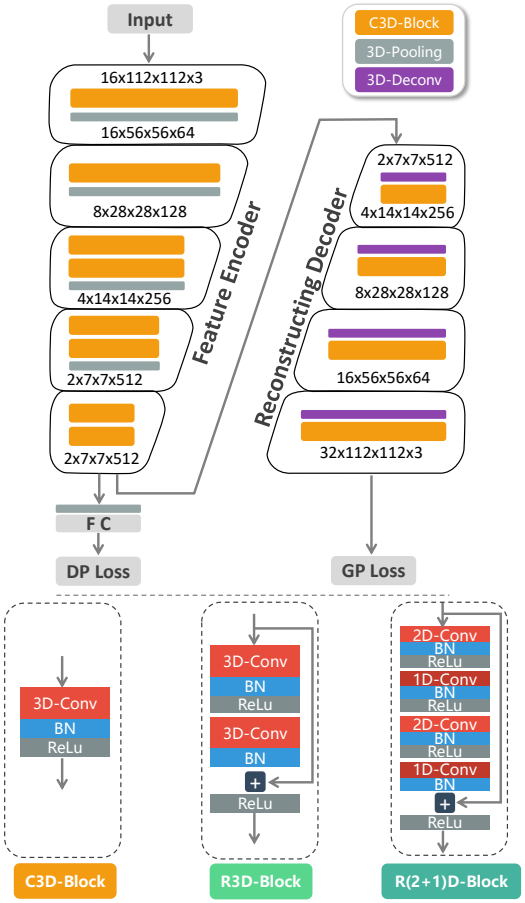

Figure 3. Up: encoder-decoder structure. Down: C3D, R3D, and R(2+1)D blocks.

CNNs for spatio-temporal representation learning as it can model the temporal information of videos. It stacks five C3D blocks which consist of a classic 3D convolution with the kernel size of $t \times k \times k$ followed by a batch normalization layer and a ReLU layer. As shown in Fig. 3, we take C3D backbone as an example to build the feature encoder and show the dimensional transformation of each block.

R3D refers to 3D CNNs with residual connections. As shown in Fig. 3, R3D block consists of two 3D convolution followed by batch normalization and ReLU layers. The input and output are connected with a residual unit before the last ReLU layer. In R(2+1)D, the overall structure is similar tp R3D. The 3D convolution is decomposed into a spatial 2D convolution and a temporal 1D convolution with additional batch normalization and ReLU layers attached.

Discriminative Perception. As shown in Fig. 3, features of the input video clip extracted by the encoder is fed to a classification model to predict the playback rate. The ground-truth label is denoted as s_c, where $1 \leq c \leq C$, C is the number of different sampling intervals of the inputs. This procedure can be referred to as discriminative perception upon a normalized probability p_c of which the input video clip belongs to class c, $p_c = \frac{\exp(a_c)}{\sum_{c=1}^{C} \exp(a_c)}$, where a_c is the c-th output of the fully connected layer. Based on the normalized probability, the parameter θ for the network model is updated by optimizing a cross entropy loss, as

$$\arg\min_{\theta} \mathcal{L}_d = -\sum_{c}^{C} s_c \log p_c. \tag{1}$$

To optimize Eq. 1, the feature encoder is driven to perceive subtle differences of motion intensity and scenario dynamics among adjacent frames which is essential for precise spatio-temporal representation.

Generative Perception. Beyond discriminative perception we further propose a generative perception mode to promote PRP's understanding capacity, which targets at reconstructing the $r\times$ slow-down video clips. The reconstruction procedure is performed with a feature decoder network which has four 3D deconvolutional blocks, Fig. 3. For each decoder block, we stack a deconvolutional layer with stride $2\times2\times2$ followed by a C3D block. To generate a video with reconstructing rate r (r times as slow as the input video), the fourth deconvolutional takes a stride of $r \times 2 \times 2$.

Ground-Truth. To predict the interpolated frames, we set the dilated sampling interval as $s = 2^{k_1}, (k_1 = 0, 1, 2, \cdots)$ and the reconstructing rate as $r = 2^{k_2}, k_2 \in 0, 1, 2$. The ground-truth of the input clip $X(2^{k_1})$ with $2^{k_2}\times$ slow-down generation can be sampled from the video $V(2^{k_1-k2})$. As shown in Fig. 2(right), a $2\times$ slow-down generative perception is implemented by taking the $2\times$ dilated sampled video clip as input and the raw video as output (self-supervision signal). If $k_2 > k_1$, we can use linear interpolation to generate the ground-truth clip from the raw video.

Motion Attention. To reconstruct video clips, MSE [14] loss is commonly used to build a generative network. It is important to note that our PRP is not designed to generate high quality videos but to learn long-short term video representations. To fulfill this purpose, we propose a motion attention regularized MSE (m-MSE) loss, which drives the network concentrating on reconstructing and interpolating frame regions in significant motion.

Denoting the t-th ground-truth frame for slow-down generation, the t-th motion attention map and the t-th predicted video frame as $G^t = (g_{ij}^t)$, $M^t = (m_{ij}^t)$ and $Y^t = (y_{ij}^t)$, the m-MSE loss can be defined as

$$\arg\min_{\theta} \mathcal{L}_g = \frac{1}{N} \sum_{t,i,j} m_{ij}^t (y_{ij}^t - g_{ij}^t)^2, \tag{2}$$

where N is the number of pixels in the predicted video clip. (ij) denotes a spatial location on video frames.

As shown in Fig. 4, the motion attention maps M are calculated according to the raw video frames $X(1)$ (de-

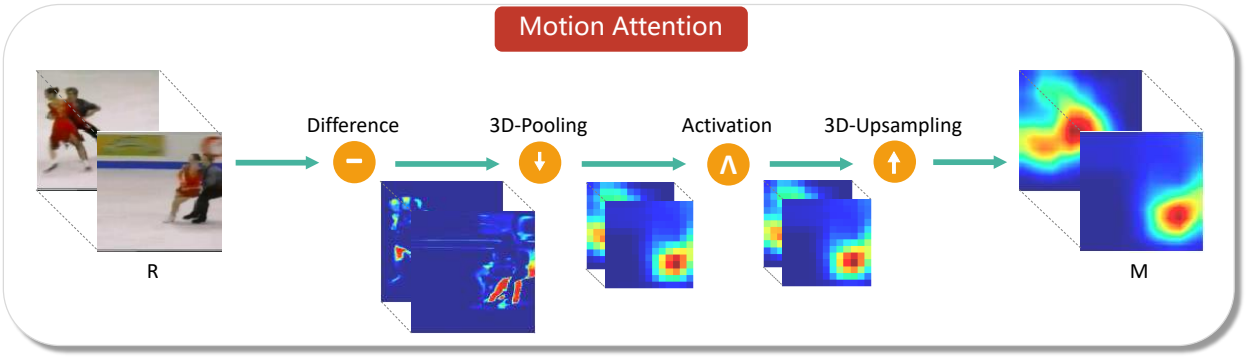

Figure 4. Calculation of motion attention based on frame difference, 3D-Pooling, activation and 3D-Upsampling operations.

noted as R) which is an $s\times$ slow-down video clip of input $X(s)$, and through four steps including difference, 3D-Pooling, activation and 3D-Upsampling. In the difference step, adjacent frames R^t and R^{t+1} from the raw video clip are used to calculate the t-th frame difference map D^t as $D^t = \mathcal{D}(R^t, R^{t+1}) = |R^t - R^{t+1}|^2$. Considering that the frame difference maps can be affected by accidental noise as well as missing static foregrounds, a 3D-Pooling operation $\mathcal{P}$, as a spatio-temporal filer, is conducted on the difference maps to make it more consistent with foregrounds and more stable in the spatio-temporal domain. Then, an increasing activation function $\mathcal{A}$ is used to transform the pixel value of the difference maps to $[\lambda_1, \lambda_2]$, $0 \leq \lambda_1 \leq 1$ and $1 \leq \lambda_2$. Finally, a 3D-Upsampling operation $\mathcal{U}$ is applied to obtain motion attention maps of the same size with the ground-truth video frames. The overall process of motion attention map generation is formulated as

$$M = \mathcal{M}(R) = \mathcal{U}(\mathcal{A}(\mathcal{P}(\mathcal{D}(R)))). \tag{3}$$

Discriminative-Generative Perception. To further learn richer spatio-temporal representations, discriminative and generative perception models are fused, Fig. 2, by optimizing the following objective function, as

$$\arg\min_{\theta} \lambda_d \mathcal{L}_d + \lambda_g \mathcal{L}_g. \tag{4}$$

Fusion is performed in a cooperative manner, as the classification model is good at identifying long-term representation for playback rate discrimination, while the generative model can capture short-term fine-details for content reconstruction. With end-to-end learning, Fig. 2, spatio-temporal characteristics of multiple temporal resolution can be encoded within the model.

3.3. Discussion

The proposed encoder-decoder framework contributes a new feature learning strategy, which is neither identical to Variational Auto-Encoder (VAE) [21] nor to Generative Adversarial Network (GAN) [41], Fig. 5. Specifically, our framework is driven by discriminative and generative models to achieve semantic retention, which means that the encoded temporal semantics can be transferred to downstream target tasks, as much as possible. By contrast, VAE targets at semantic sampling controlled by the latent variable (z) following normal distribution. The encoder in VAE should learn features that best represent the distribution of inputs while the generator uses specified features for data generation conditioned on the latent variable.

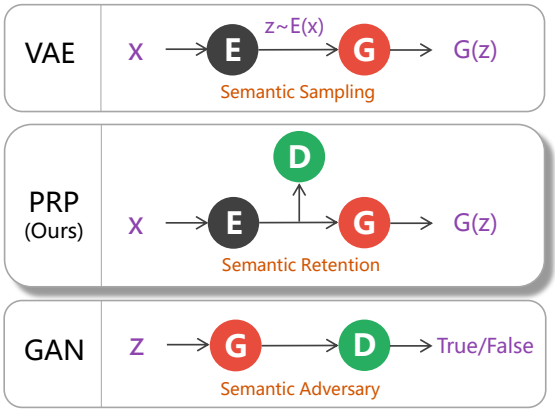

Figure 5. Comparison of Variational Auto-Encoder (VAE), Generative Adversarial Network (GAN), and the proposed encoder-decoder model. 'E', 'D', and 'G' denote 'Encoder', 'Discriminator' and 'Generator', respectively.

Like GAN, our approach involves both generative and discriminative models. The essential difference is that GAN leverages models in an adversarial manner while ours works cooperatively. GAN uses the generative model to produce images which are difficult to be classified by the discriminative model. Our approach learns general semantics, $i.e.$, multi-resolution spatio-temporal representation, in a cooperative discrimination-generation manner.

4. Experiments

We first elaborate experimental settings for PRP, and then evaluate various sampling intervals and reconstructing rates with ablation study on a target task (action recognition). We then analyze how PRP drives the model focusing on foreground regions and perceiving long-short term spatio-temporal characteristics. Finally, we evaluate the performance of PRP by applying the self-supervised models on target tasks including video action recognition and video retrieval, and compare it with state-of-the-art methods.

4.1. Experimental Setting

Datasets. Two action recognition datasets, UCF101 [31] and HMDB51 [15], are used to demonstrate the effectiveness of PRP. UCF101 is collected from websites including Prelinger archive, YouTube and Google videos, containing 101 action categories with 9.5k videos for training and 3.5k videos for testing. HMDB51 is extracted from a variety of sources ranging from digitized movies to YouTube. It is consists of 51 action categories with 3.4k videos for training and 1.4k videos for testing. Both datasets exhibit challenges include intra-class variance of actions, complex camera motions, and cluttered backgrounds. To perform action recognition and retrieval on these datasets requires learning rich spatio-temporal representation.

Network Architecture. In video encoder, C3D, R3D, R(2+1)D are used as network backbones, where the kernel size of 3D convolutional layers is set to $3 \times 3 \times 3$. In video generation, four deconvolutional layers are stacked and followed by C3D blocks. To generate a video which is r times as slow as the input video, we set the $4\text{-}th$ deconvolutional layer with a stride of $r \times 2 \times 2$, where the reconstructing rate r is determined through ablation study.

Motion Activation. To calculate motion attention maps, the activation function $\mathcal{A}$ in Eq. 3 is implemented as $\mathcal{A}(D) = \frac{\lambda_2 - \lambda_1}{max(D) - min(D)}(D - min(D)) + \lambda_1$, where D is the frame difference map. λ_1 is empirically set to 0.8 and λ_2 2.0. We use an 3D-AveragePooling with kernel size $15 \times 28 \times 28$ and stride size $16 \times 7 \times 7$. The 3D-Upsampling operation is set to tri-linear mode.

Parameters. Following the settings in [32, 33], we set the length of input video $X(s)$ $l = 16$ and determine the dilated sampling interval $s \in S$ through ablation study. During training, we randomly split 800 videos from the training set as validation set. Video frames are resized to 128×171 and randomly cropped to 112×112 as data augmentation. We empirically set the parameters λ_d, λ_g for loss balance as 0.1 and 1. With a initial learning rate 0.01, momentum 0.9 and weight decay 0.0005, the pre-training process is carried out for 300 epochs. The learned representation model with the lowest validation loss is used for target tasks.

Samp. Interval	Random acc.(%)	DP acc. (%)
{1,2}	50	88.3
{1,2,4}	33	80.1
{1,2,4,8}	25	69.7
{1,2,4,8,16}	20	60.1

Table 1. Classification accuracy of the discriminative perception (DP) model under different sampling intervals.

Method	Samp. Interval	Rec. Rate	UCF101(%)
Random	-	-	62.0
DP	{1,2}	-	68.3
	{1,2,4}	-	68.7
	{1,2,4,8}	-	**69.9**
	{1,2,4,8,16}	-	67.9
GP	{1,2,4,8}	1 (w/o *MA*)	67.1
	{1,2,4,8}	1 (w/ *MA*)	68.1
	{1,2,4,8}	2 (w/ *MA*)	68.2
	{1,2,4,8}	4 (w/ *MA*)	**68.4**
DG-P	{1,2,4,8}	2 (w/ *MA*)	**70.9**

Table 2. Ablation study of different model perception methods with corresponding different model parameters. The figures refer to action recognition accuracy on UCF101. "Sam.Rate" and "Rec.Rate" respectively denote sampling interval and reconstructing rate. "DP", "GP", and "DG-P" respectively denote discriminative perception, generative perception, and discriminative-generative perception. "*MA*" denotes *Motion Attention*.

4.2. Ablation study

In this section, we conduct experiments on the first split of UCF101 to analyze the effect of PRP under different dilated sampling intervals, different reconstructing rates, with/without motion attention.

Dilated sampling interval. As shown in Table 1, discriminative perception accuracy is consistently higher than random accuracy, which indicates that the discriminative perception model can learn effective spatio-temporal representation. Specifically, as the sampling interval s increases, discriminative perception accuracy gradually decreases from 88.3% to 60.1%, while the accuracy of the target task increases from 68.3% of {1,2} to 69.9% with sampling interval {1,2,4,8}, Table 2. It manifests that to some extent, larger sampling intervals force the model perceiving longer motion information which improves the representation capability of the learned model. However, when it comes to {1,2,4,8,16}, the video content jumps too much to be well perceived which makes the model confuse to learn discriminative representations. Therefore, the action recognition accuracy stops increasing. We thus set a sampling interval $s \in \{1, 2, 4, 8\}$ in the following experiments.

Reconstructing rate. As shown in Table 2, with reconstructing rate r increasing, the performance increases

from 68.1% to 68.4% when motion attention loss is applied, which can be explained that large reconstruction rate r can force the network focusing on motion details, which is helpful for video understanding. Considering the performance of $r = 2$ is comparable to which of $r = 4$, we set $r = 2$ as default in what follows to reduce the computational cost of the network.

Discriminative and Generative Perception. As shown in Table 2, discriminative perception improves the action recognition accuracy from 62.0% to 69.9%, while generative perception improves the accuracy from 62.0% to 68.4%. The discriminative-generative model further improves the accuracy to 70.9%, which validates the effectiveness of cooperative work of these two branches.

Motion Attention. The motion attention mechnism can drives representation to focus on meaningful foreground regions. As shown in Table 2, the application of motion attention boosts the accuracy from 67.1% to 68.1%, which is also a significant margin considering the challenging action recognition task.

4.3. Visualizing Self-supervised Representation

We try to understand what PRP learns by visualizing the feature activation maps, which indicating where the spatio-temporal representation focuses on. In Fig. 6, we visualize and compare different perception models' activation maps on video frames. It can be seen that the discriminative perception model (DP) learns features sensitive to incomplete foreground regions containing major motion information, while the generative perception (GP) model learns features sensitive to where motion occurs but diverse to more context regions. With motion attention preferring to enhance motion areas, the generative perception model produces activation map with more motion areas activated. By fusing these two models, the learned features focus on complete foreground regions, which implies that the representation model incorporates long-short term motion information.

4.4. Evaluating Self-supervised Representation

Action Recognition. To verify our findings, we conduct experiments on action recognition which is a representative target task to validate the effectiveness of self-supervised representation [39]. For action recognition, we initialize the backbones with the model pre-trained on the first split of UCF101 by PRP, and fine-tune on UCF101 and HMDB51, Table 3. Data pre-processing and experimental settings are the same as those during PRP training. We feed features extracted by the backbones to fully-connected layers and obtain the category prediction. For training, the fine-tuning procedure stops after 150 epochs. For testing, we follow the protocol of [33] and sample 10 clips for each video. The predictions on the sampled clips are then averaged to obtain the final prediction results. And we average classification accuracy over 3 splits for fair comparison.

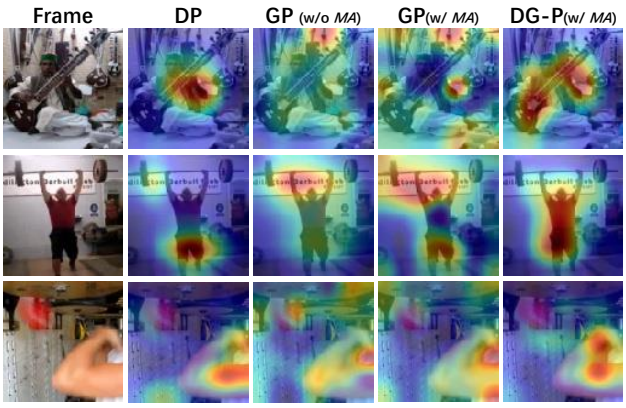

Figure 6. Visualization of activation maps.The attention maps are generated by summarizing convolutional feature channels in conv5 layer [40] ."DP", "GP", and "DG-P" respectively denote discriminative perception, generative perception, and discriminative-generative perception. "*MA*" denotes *Motion Attention*.

Method	UCF101(%)	HMDB51(%)
Jigsaw[28]	51.5	22.5
OPN[24]	56.3	22.1
Büchler[1]	58.6	25.0
Mas[36]	58.8	32.6
3D ST-puzzle[19]	65.0	31.3
ImageNet pre-trained	67.1	28.5
C3D(random)	61.8	24.7
C3D(VCOP[39])	65.6	28.4
C3D(PRP)	**69.1**	**34.5**
R3D(random)	54.5	23.4
R3D (VCOP[39])	64.9	29.5
R3D (PRP)	**66.5**	**29.7**
R(2+1)D(random)	55.8	22.0
R(2+1)D(VCOP[39])	**72.4**	30.9
R(2+1)D(PRP)	72.1	**35.0**

Table 3. Performance comparison of self-supervised methods for spatio-temporal representation learning on UCF101 and HMDB51.

With the C3D backbone, our PRP approach obtains 69.1% and 34.5% which is 7.3% and 9.8% better than random initialization on UCF101 and HMDB51 respectively, Table 3. Our PRP approach also obtains 3.5% and 6.1% better results compared with state-of-the-art VCOP approach [39], which are significant margins for the challenging action recognition task. With the R(2+1) backbone, PRP achieves 16.3% (72.1% vs. 55.8%) and 13.0% (35.0% vs. 22.0%) improvement over the random initialization. Our PRP approach also outperforms VCOP with significant margins. With the obtained results, we validate that PRP is able to learn richer spatio-temporal representations of videos compared with previous methods.

Figure 7. Comparison of video retrieval results. Red fonts indicate correct retrieval results. It can be seen that compared with the state-of-the-art VCOP approach, PRP achieves more accurate and reasonable video retrieval results. (Best viewed in color)

Methods	top1	top5	top10	top20	top50
Jigsaw[28]	19.7	28.5	33.5	40.0	49.4
OPN[24]	19.9	28.7	34.0	40.6	51.6
Büchler[1]	25.7	36.2	42.2	49.2	59.5
C3D(random)	16.7	27.5	33.7	41.4	53.0
C3D(VCOP[39])	12.5	29.0	39.0	50.6	66.9
C3D(PRP)	**23.2**	**38.1**	**46.0**	**55.7**	**68.4**
R3D(random)	9.9	18.9	26.0	35.5	51.9
R3D(VCOP[39])	14.1	30.3	40.4	51.1	66.5
R3D(PRP)	**22.8**	**38.5**	**46.7**	**55.2**	**69.1**
R(2+1)D(random)	10.6	20.7	27.4	37.4	53.1
R(2+1)D(VCOP[39])	10.7	25.9	35.4	47.3	63.9
R(2+1)D(PRP)	**20.3**	**34.0**	**41.9**	**51.7**	**64.2**

Table 4. Video retrieval performance on UCF101.

Video Retrieval. To further verify its effectiveness, PRP is tested on the target task of nearest-neighbor video retrieval. As the video retrieval task is conducted with features extracted by the backbone network without fine-tuning, they largely rely upon the representative capacity of self-supervised model. An experiment is carried out on the first split of UCF101, following the protocol in [39]. In the process of retrieval, video convolutional features are extracted with the backbone pre-trained by PRP. Each video in the test set is used to query k nearest videos from the training set based upon their spatio-temporal features. When the category in the retrieved result is identical to that in the test video, we count this as the correct retrieval.

In Table 4 and Table 5, we show top-1, top-5, top-10, top-20, and top-50 retrieval accuracy, which shows that PRP outperforms the state-of-the-art method equivalent on all evaluation metrics by substantial margins (8.7∼10.7% for top1 accuracy on UCF101). In Fig. 7, qualitative results further show PRP's superiority.

Methods	top1	top5	top10	top20	top50
C3D(random)	7.4	20.5	31.9	44.5	66.3
C3D(VCOP[39])	7.4	22.6	34.4	48.5	70.1
C3D(PRP)	**10.5**	**27.2**	**40.4**	**56.2**	**75.9**
R3D(random)	6.7	18.3	28.3	43.1	67.9
R3D(VCOP[39])	7.6	22.9	34.4	48.8	68.9
R3D(PRP)	**8.2**	**25.8**	**38.5**	**53.3**	**75.9**
R(2+1)D(random)	4.5	14.8	23.4	38.9	63.0
R(2+1)D(VCOP[39])	5.7	19.5	30.7	45.8	67.0
R(2+1)D(PRP)	**8.2**	**25.3**	**36.2**	**51.0**	**73.0**

Table 5. Video retrieval performance on HMDB51.

5. Conclusion

In this paper, we proposed a novel video Playback Rate Perception (PRP) approach for self-supervised spatio-temporal representation learning. With a simple dilated sampling strategy, we augmented videos into different temporal-resolutions, which were then used to learn the long-short term characteristics of videos with discriminative and generative models. Self-supervised models were applied on video action recognition and video retrieval tasks. Extensive experiments showed that self-supervised models, trained with PRP, outperformed state-of-the-art self-supervised models with significant margins. Our work presented a promising direction and a new framework for self-supervised spatio-temporal representation learning.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61836012, 61671427 and 61771447, and the National Key RD Program of China (2017YFB1002400).

References

[1] Uta Buchler, Biagio Brattoli, and Bjorn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In *ECCV*, pages 770–786, 2018. 7, 8

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

[3] Pathak Deepak, B. Girshick Ross, Dollár Piotr, Darrell Trevor, and Hariharan Bharath. Learning features by watching objects move. In *IEEE CVPR*, pages 6024–6033, 2017. 2

[4] P. Kingma Diederik and Welling Max. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[5] Jayaraman Dinesh and Grauman Kristen. Learning image representations tied to egomotion from unlabeled video. *Int. J. Com. Vis.*, 125(1-3):136–161, 2017. 2

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

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE ICCV*, pages 6202–6211, 2019. 2

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE CVPR*, pages 1933–1941, 2016. 3

[9] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *IEEE CVPR*, pages 10364–10374, 2019. 2

[10] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *IEEE CVPR*, pages 3636–3645, 2017. 1, 2

[11] MChuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *IEEE ICCV*, pages 7053–7062, 2019. 2

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[13] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *IEEE ICCV*, pages 6391–6400, 2019. 2

[14] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *IEEE CVPR*, pages 733–742, 2016. 4

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

[16] Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Li Kai, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009. 1

[17] S. J. Ren Jimmy, Hu Yongtao, Tai Yu-Wing, Wang Chuan, Xu Li, Sun Wenxiu, and Yan Qiong. Look, listen and learn-A multimodal LSTM for speaker identification. In *AAAI*, pages 3581–3587, 2016. 2

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[19] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, volume 33, pages 8545–8552, 2019. 1, 3, 7

[20] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *WACV*, pages 793–802. IEEE, 2018. 2

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1106–1114, 2012. 1

[23] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *IEEE CVPR*, pages 6874–6883, 2017. 2

[24] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *IEEE ICCV*, pages 667–676, 2017. 2, 7, 8

[25] Margaret Livingstone and David Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240(4853):740–749, 1988. 2

[26] Margaret Livingstone and David Hubel. Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13(1):1–10, 1994. 2

[27] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544. Springer, 2016. 2

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 7, 8

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

[30] Arandjelovic Relja and Zisserman Andrew. Look, listen and learn. In *IEEE ICCV*, pages 609–617, 2017. 2

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[32] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE CVPR*, pages 4489–4497, 2015. 3, 6

[33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE CVPR*, pages 6450–6459, 2018. 3, 6, 7

[34] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, pages 613–621, 2016. 2

[35] Fang Wan, Pengxu Wei, Zhenjun Han, Jianbin Jiao, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(10):2395–2409, 2019. 1

[36] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *IEEE CVPR*, pages 4006–4015, 2019. 2, 3, 7

[37] Wang Xiaolong and Gupta Abhinav. Unsupervised learning of visual representations using videos. In *IEEE CVPR*, pages 2794–2802, 2015. 2

[38] Wang Xiaolong, He Kaiming, and Gupta Abhinav. Transitive invariance for self-supervised visual representation learning. In *IEEE CVPR*, pages 1338–1347, 2017. 2

[39] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *IEEE CVPR*, pages 10334–10343, 2019. 3, 7, 8

[40] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 7

[41] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, pages 7354–7363, 2019. 5

[42] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *NeurIPS*, pages 147–155, 2019. 2

[43] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *ACM Multimedia*, pages 1933–1941, 2017. 2