BRINGING SACCADES AND FIXATIONS INTO SELF-SUPERVISED VIDEO REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a self-supervised video representation learning (video SSL) method by taking inspiration from cognitive science and neuroscience on human visual perception. Different from previous methods that focus on the inherent properties of videos, we argue that humans learn to perceive the world through the self-awareness of the semantic changes or consistency in the input stimuli in the absence of labels, accompanied by representation reorganization during the post-learning rest periods. To this end, we first exploit the presence of saccades as an indicator of semantic changes in a contrastive learning framework, mimicking the self-awareness in human representation learning. The saccades are generated artificially without eye-tracking data. Second, we model the semantic consistency in eye fixation by minimizing the prediction error between the predicted and the true state of another time point. Finally, we incorporate prototypical contrastive learning to reorganize the learned representations to enhance the associations among perceptually similar ones. Compared to previous video SSL solutions, our method can capture finer-grained semantics from video instances, and the associations among similar ones are further strengthened. Experiments show that the proposed bio-inspired video SSL method significantly improves the Top-1 video retrieval accuracy on UCF101 and achieves superior performance on downstream tasks such as action recognition under comparable settings.

1 INTRODUCTION

Learning without labels is the most common way for humans to get to know the world (DiCarlo et al., 2012), and it has also been widely studied in machine learning for developing intelligent agents. In particular, many researchers focus on self-supervised learning (SSL) from dynamic visual input data, *i.e.*, videos (Hurri & Hyvärinen, 2003; Mobahi et al., 2009; Srivastava et al., 2015), which comes closest to the natural data perceived by humans. Recently, deep learning based video SSL methods have also shown superior performance over traditional non-deep learning methods (Wang et al., 2021a; Duan et al., 2022). However, there is still a large room for improvement considering the gap between the unsupervised learning abilities of deep learning models and humans.

One notable difference between deep video SSL methods and human unsupervised learning is that the former typically learn discriminative representations by considering the inherent data properties, such as the clip order (Misra et al., 2016), the spatiotemporal coherence (Vondrick et al., 2018; Ding et al., 2022; Vondrick et al., 2016), the transformations exerted on the video (Jenni et al., 2020), the intrinsic speed of the video (Benaim et al., 2020), *etc*., and propose various pretext tasks accordingly. While for humans, the self-awareness of the **semantic change or consistency** in the input stimuli is essential for learning without labels (Melcher & Colby, 2008). Besides, the encoded representations in the brain are not left unchanged but kept being **reorganized** to yield a representation structure with strengthened associations among perceptually similar representations (Diekelmann & Born, 2010). Fig. 1 shows an overall comparison. This discrepancy inspired us to propose a new video SSL method by taking inspiration from cognitive science and neuroscience on human visual perception. Recently, Illing et al. (2021) proposed a biologically plausible unsupervised learning rule that treats the presence of saccades as a global synaptic modulator. However, it is less powerful due to the inherent difficulty in optimizing deep networks with layer-wise optimization.

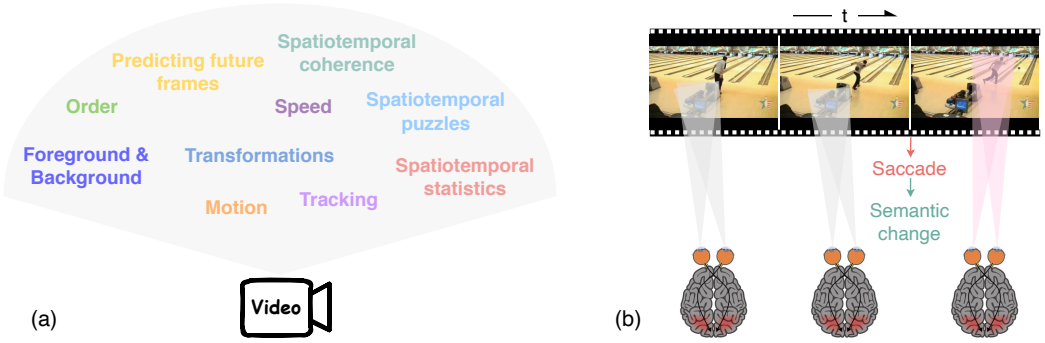

Figure 1: **Overall comparison. (a) Previous video SSL methods** design pretext tasks based on the inherent properties of videos, while **(b) our method** explores the presence of saccades as an indicator of semantic change to mimic the role of self-awareness in human perception.

Human visual perception is mainly accomplished by alternating *saccade* and *fixation* when the heads are relatively still. The former is the rapid foveal motion from one target of interest to another, while the latter is the period where the eye is kept aligned with one target for the processing of the visual details. To capture the **semantic change** in the video, we propose to exploit the presence of saccades as an indicator of the semantic change and propose a semantic-change-aware contrastive learning framework. This is inspired by the fact that the human would perform a saccade when a semantic change occurs in the visual inputs. Specifically, the positive pairs are formed by features of the same fixation location in a video, and the negative pairs are formed by features of different fixation locations in the same video or features from different videos. Compared to previous contrastive-based video SSL methods, our method makes distinctions within the same video indicated by saccades and achieves finer-grained semantic learning. Note that we mimic saccades and fixations by exerting different fixation masks on the input without eye-tracking data, making this SSL scheme a general one for any video data without extra supervision.

We further encourage the **semantic consistency** within a fixation duration by minimizing the prediction error (PE) when using the current state to predict that of another time point in the same fixation duration. In this way, PE can serve as an extra supervision signal to avoid semantic discrepancy during semantic-change-aware contrastive learning. This is also biorational, as PE is known as an important modulator in perception, attention, and motivation control (Den Ouden et al., 2012).

To enhance the association among the previously learned finer-grained semantics, inspired by the **reorganization** in human representation learning (Diekelmann & Born, 2010), we incorporate prototypical contrastive learning (Li et al., 2020) to gradually redistribute the representations. The learned representations are pulled towards their corresponding prototypes and pushed away from other prototypes. Such post-learning reorganization facilitates grouping unseen input stimuli into meaningful categories based on similarity, which leads to improved Top-1 retrieval accuracy compared with previous contrastive-based video SSL methods.

In summary, we propose a video SSL framework by taking inspiration from cognitive science and neuroscience on human visual perception. We first exploit the presence of saccades as an indicator of semantic change in a contrastive learning framework for modeling the role of self-awareness in human representation learning. Then, we model the semantic consistency in the input by minimizing PE between a predicted and the true states of different time points during a fixation. Third, we incorporate prototypical contrastive learning to reorganize the learned representations such that the associations among perceptually similar ones would be strengthened after redistribution. Experiments show that the proposed bio-inspired video SSL method significantly improves the Top-1 video retrieval accuracy on UCF101, and achieves superior performance on downstream tasks such as action recognition. The code and the pre-trained models will be released.

2 Related Work

Self-supervised learning has been studied with various data formats, including image (Wu et al., 2018; Grill et al., 2020; He et al., 2020; Li et al., 2020), video (Xu et al., 2019; Benaim et al., 2020;

Qian et al., 2021a; Duan et al., 2022), and multi-modal data (Alayrac et al., 2020; Patrick et al., 2021). In this section, we focus on self-supervised representation learning on videos.

Non-contrastive video SSL methods. Previous video SSL methods mainly learn discriminative representations by designing various *pretext tasks* based on the analysis of the inherent spatiotemporal properties of video data. The pretext tasks include figuring out the correct order of the clips (Misra et al., 2016; Fernando et al., 2017; Lee et al., 2017; Wei et al., 2018; Xu et al., 2019), tracking contents across adjacent frames (Wang & Gupta, 2015; Vondrick et al., 2016; Pathak et al., 2017; Vondrick et al., 2018; Wang et al., 2019b), studying foreground and background robustness (Luo et al., 2017; Wang et al., 2021b;c; Ding et al., 2022), predicting future frames (Vondrick et al., 2016; Luo et al., 2017; Villegas et al., 2017; Han et al., 2020a; Behrmann et al., 2021), solving spatiotemporal puzzles (Kim et al., 2019) or video cloze (Luo et al., 2020), learning the spatiotemporal statistics of the videos (Wang et al., 2019a; 2021a), recognizing transformations exerted on the video (Jenni et al., 2020; Duan et al., 2022) , or determining whether a video is played at the intrinsic speeds (Benaim et al., 2020; Wang et al., 2020; Yao et al., 2020; Chen et al., 2021).

Contrastive video SSL methods. Another line of research adapts spatiotemporal properties into the *contrastive learning* framework (Hadsell et al., 2006) by constructing the positive and negative pairs based on various spatiotemporal cues. More specifically, some methods extend the instance discrimination methods in image SSL (Wu et al., 2018), and directly use clips randomly sampled from the same video as a positive pair (Han et al., 2020b; Lin et al., 2021; Pan et al., 2021; Qian et al., 2021b; Yao et al., 2021). Some methods consider the spatiotemporal consistency of video and treat the predicted and the ground-truth features at the same spatiotemporal location as a positive pair (Han et al., 2019; 2020a). Some methods relate video understanding to pace reasoning ability and construct positive pairs by sampling clips with different sampling rates from the same video or sampling clips with the same pace from different videos (Huang et al., 2021). Some methods construct the positive pair from both frame-level and video-level representations (Kong et al., 2020; Kuang et al., 2021). Some others construct positive pairs through spatiotemporal data augmentations (Qian et al., 2021b; Sun et al., 2021), or exploit motion information for data augmentation (Dwibedi et al., 2019; Li et al., 2021; Wang et al., 2021b). Our method also belongs to this line of research. However, we construct positive and negative pairs based on the semantic change indicated by saccades, resulting in a finer-grained distinction of the semantics from the same video instance.

Comparison with CLAPP. Both the CLAPP model (Illing et al., 2021) and our work exploit the self-awareness of saccades for self-supervised representation learning, but we differ significantly in several aspects. First, CLAPP aims to propose a local learning rule for building deep representations without back-propagation, where each layer is trained independently to predict whether a saccade happens. While ours utilizes the presence of saccades to construct the positive and negative pairs in a contrastive learning framework for end-to-end learning. Second, for each layer, CLAPP restricts the predictions to be similar to its responses to future inputs and as different as possible from its responses to fake inputs. However, in our work, we consider minimizing the discrepancies between the predictions and the future responses in the absence of a saccade for semantic consistency modeling. Third, besides constructing a saccade by switching the network inputs from one video to another as in CLAPP, we also consider inter-video saccades, which are realized by intentionally changing the fixation area on the same video to capture finer-grained semantics. As shown in §4, we significantly improve over CLAPP for the video recognition task on UCF101.

3 METHOD

As shown in Fig. 2, our method consists of three parts. First, we explore saccades as the indicator of semantic change in a contrastive learning process (§3.1), where the negative pairs consist of features before and after a saccade occurs. Second, we model the semantic consistency during fixation by minimizing the prediction error (PE) between the predicted and the true features of different time points (§3.2). Third, we perform a post-learning reorganization to strengthen the associations among perceptually similar representations (§3.3).

3.1 CAPTURING SEMANTIC CHANGE VIA CONTRASTIVE LEARNING

Preparing saccades. As the ground-truth gaze data are generally collected using eye-tracking devices, which requires a lot of manual effort, and usually has personal heterogeneity considering the exact fixation locations, in our design, we consider a restricted number of coarse fixation lo-

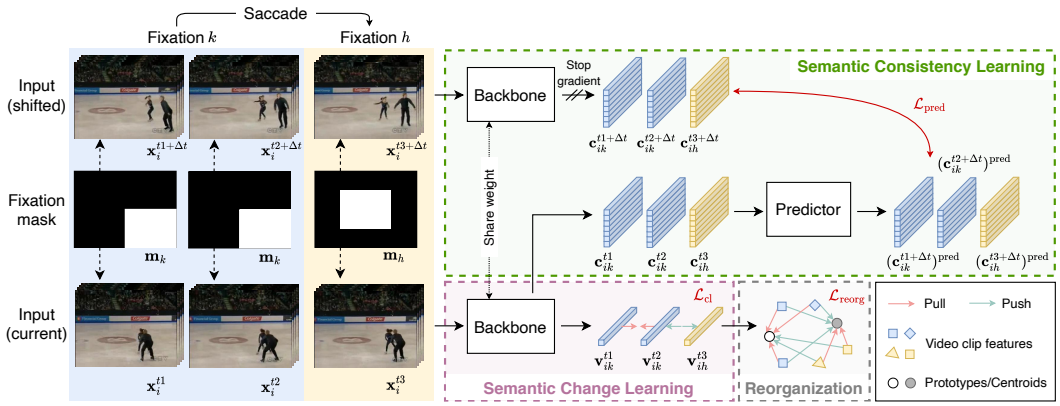

Figure 2: **Framework of the proposed bio-inspired video SSL method.** Given a video $\mathbf{x}_i$, the positive and negative pairs for semantic-change-aware contrastive learning can be constructed by exerting the same or different fixation masks $\mathbf{m}$ on different clips of the video (§3.1), respectively. Further, the semantic consistency is modeled by minimizing the prediction error between the predicted state $(\mathbf{c}_{ik}^{t*+\Delta t})^{\text{pred}}$ and the true state of another time point $\mathbf{c}_{ik}^{t*+\Delta t}$ (§3.2). Finally, the learned representations are reorganized through a prototypical contrastive learning process to strengthen the association among perceptually similar ones (§3.3).

cations, and "creating" saccades by intentionally changing the fixation locations and exerting the corresponding fixation masks $\{\mathbf{m}_i\}_{i=1}^{N_p}$ on the input stimuli. To balance the performance and efficiency, we set $N_p = 5$, *i.e.*, we select 5 fixation locations and generate 5 masks accordingly to simulate the *pupil function* of an ideal system, where the value is 1 at every point within the fixation area and 0 otherwise. Fixation masks used in our experiments are shown in Fig. 3 (a).

Semantic-change-aware contrastive learning. To capture the semantic change in the presence of a saccade, we try to make the representations before and after a saccade distinct from each other. To this end, we treat such representations as negative pairs and otherwise positive pairs and optimize the latent feature space by minimizing the contrastive loss.

Formally, given a training set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of N videos without category labels, we aim to learn an embedding function f_θ to map $\mathbf{X}$ to features $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$, where $\mathbf{v}_i = f_\theta(\mathbf{x}_i)$ and $\mathbf{v}_i \in \mathbb{R}^D$ is expected to capture the semantics of $\mathbf{x}_i$. The semantic-change-aware contrastive loss is inspired by InfoNCE (Oord et al., 2018; He et al., 2020), and is calculated as follows:

$$\mathcal{L}_{\text{cl}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(\mathbf{v}_{ik}\cdot\mathbf{v}'_{ik}/\tau)}{\exp(\mathbf{v}_{ik}\cdot\mathbf{v}'_{ik}/\tau) + \sum_{j\in\mathcal{I}_-}\exp(\mathbf{v}_{ik}\cdot\mathbf{v}'_{jh}/\tau))}, \quad |\mathcal{I}_-| = N_{\text{neg}}. \tag{1}$$

Here, $\mathbf{v}'_{ik} = f_\theta(\mathbf{x}'_i \odot \mathbf{m}_k)$ is a positive sample for $\mathbf{v}_{ik} = f_\theta(\mathbf{x}_i \odot \mathbf{m}_k)$ ($\odot$ is Hadamard product), $\mathbf{x}'_i$ is an augmented version of $\mathbf{x}_i$, $\mathbf{v}'_{jh} \neq \mathbf{v}'_{ik}$ is a negative sample, $\mathcal{I}_-$ is the set of indices for N_{neg} selected negative samples, and $k, h \in \{1, 2, \ldots, N_p\}$ are indices of the mask $\mathbf{m}$. Note that a sample is considered positive if and only if $j = i$ and $h = k$, *i.e.*, they are from the same fixation region of the same video. By minimizing Eq. (1), the embedding function f_θ is trained to distinguish between finer-grained semantics within a video. Perceptually similar finer-grained semantics will be further associated together through a reorganization process.

Memory bank. As previously revealed in (Oord et al., 2018; Wu et al., 2018), a large number of negative pairs is essential for training InfoNCE loss, which is typically restricted by the batch size. To alleviate this issue, we follow (Wu et al., 2018) and maintain a memory bank $\overline{\mathbf{V}} = \{\overline{\mathbf{v}}_i\}_{i=1}^{N*N_p}$ for different fixation locations of all the videos in the training dataset. Here, N is the number of training videos, and N_p is the number of masks. Similar to (Wu et al., 2018), we initialize $\overline{\mathbf{V}}$ with random D-dimensional unit vectors and update the slot $\overline{\mathbf{v}}_i$ with the latest feature $\mathbf{v}_i$ as follows:

$$\overline{\mathbf{v}}_i \leftarrow (1-m)\overline{\mathbf{v}}_i + m\mathbf{v}_i, \tag{2}$$

where $m \in [0, 1]$ is a momentum value. With $\overline{\mathbf{V}}$, we can rewrite the contrastive learning procedure and Eq. (1) by replacing the negative samples $\mathbf{v}'$ by their memory bank representations $\overline{\mathbf{v}}$.

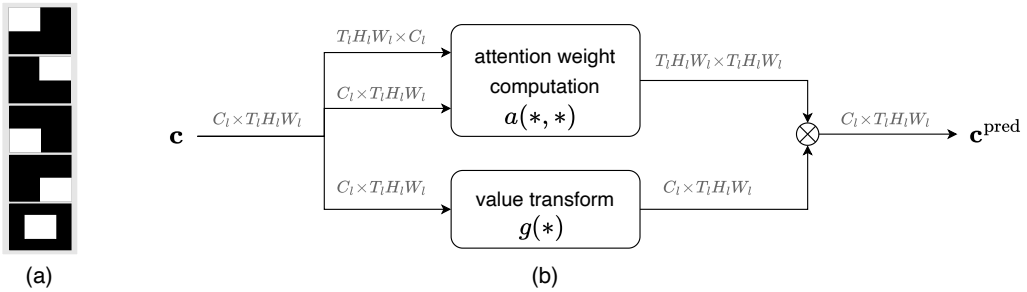

Figure 3: Method details. **(a) Exemplar fixation masks** for $N_p = 5$, which cover the major region in the visual field. See §3.1) for more details. **(b) Prediction module** introduced in §3.2), which is designed as a variant of the self-attention module to capture the spatiotemporal correlations.

3.2 Modeling semantic consistency via minimizing PE

To encourage the semantic consistency between the states of two time points within the a fixation, given a video $\mathbf{x}_i$, we minimize the PE when using $\mathbf{c}_i^t$ to predict $\mathbf{c}_i^{t+\Delta t}$, where $\mathbf{c}_i^* \in \mathbb{R}^{C_l \times (T_l H_l W_l)}$ is reshaped from the l-th level feature from the backbone f_θ whose size is $C_l \times T_l \times H_l \times W_l$, and C_l, T_l, H_l, W_l are the feature channel number, the temporal resolution, the height, and the width, respectively. Note that t is the time point of a clip from $\mathbf{x}_i$, and Δt can be either positive or negative.

The predictor module p_θ is designed as a variant of the self-attention module (Bahdanau et al., 2015; Xie et al., 2021) to capture the spatiotemporal correlations, which first obtains the spatiotemporal weights based on the element-wise cosine similarity and then calculates each predicted element as the weighted sum of the transformed version of all other spatiotemporal elements. An illustration is shown in Fig. 3 (b). The u-th element of the predicted feature $(\mathbf{c}_i^{t+\Delta t})^{\text{pred}}$ is calculated as:

$$(\mathbf{c}_{iu}^{t+\Delta t})^{\text{pred}} = \sum_{v=1}^{T_l H_l W_l} a(\mathbf{c}_{iu}^t, \mathbf{c}_{iv}^t) \cdot g(\mathbf{c}_{iv}^t), \tag{3}$$

where $a(\cdot,\cdot)$ calculates the attention weights as:

$$a(\mathbf{c}_{iu}^t, \mathbf{c}_{iv}^t) = \text{ReLU}(\cos(\mathbf{c}_{iu}^t, \mathbf{c}_{iv}^t)), \tag{4}$$

and the transform function $g(\cdot)$ is a linear layer that maps a C_l-dim input to a C_l-dim output. The semantic consistency is optimized by minimizing the prediction loss defined as follows:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_l H_l W_l} \sum_{u=1}^{T_l H_l W_l} |\mathbf{c}_{iu}^{t+\Delta t} - (\mathbf{c}_{iu}^{t+\Delta t})^{\text{pred}}|. \tag{5}$$

3.3 Reorganizing via prototypical contrastive learning

Inspired by the gradual redistribution and reorganization of memory representations during the post-learning rest periods (Diekelmann & Born, 2010), after training the whole framework with $\mathcal{L}_{\text{cl}}$ (Eq. (1)) and $\mathcal{L}_{\text{pred}}$ (Eq. (5)) to converge, we further perform reorganization through prototypical contrastive learning (Li et al., 2020) to strengthen the associations among similar representations.

Specifically, we first cluster the representations $\{\mathbf{v}_{ik}\}$ for R times to obtain R distinct clustering results $\{\mathbf{G}^{(r)}\}_{r=1}^{R}$, where $\mathbf{G}^{(r)}$ contains $Q^{(r)}$ clusters. Then, we randomly pick $Q = \min\{Q^{(r)}, N_{\text{neg}}\}$ clusters from each $\mathbf{G}^{(r)}$ to form $\mathbf{G}^{(r)\prime} = \{G_q^{(r)}\}_{q=1}^{Q}$, where the centroid of $G_q^{(r)}$ is $\mathbf{o}_q^{(r)}$. The reorganization loss is calculated as follows:

$$\mathcal{L}_{\text{reorg}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N_p} \sum_{k=1}^{N_p} \frac{1}{R} \sum_{r=1}^{R} -\log \frac{\exp(\mathbf{v}_{ik} \cdot \mathbf{o}_s^{(r)} / \phi_s^{(r)})}{\sum_{j=0}^{Q} \exp(\mathbf{v}_{ik} \cdot \mathbf{o}_j^{(r)} / \phi_j^{(r)}))}. \tag{6}$$

Here, $G_s^{(r)}$ is the cluster to which $\mathbf{v}_{ik}$ is assigned, $\mathbf{o}_s^{(r)}$ is the centroid of $G_s^{(r)}$, and $\phi_*^{(r)}$ denotes the concentration estimation of the cluster $G_*^{(r)}$ as in (Li et al., 2020). In this way, the associations among previously learned finer-grained semantics are further strengthened, which can facilitate similarity-based categorization for unknown input stimuli.

Table 1: **Video retrieval results on UCF101 and HMDB51.** All the models are pre-trained on UCF101. Larger values are better. See §4.2 for details.

	Method	Backbone	UCF101				HMDB51			
			Top-1	Top-5	Top-10	Top-20	Top-1	Top-5	Top-10	Top-20
Non-contrastive video SSL methods	VCOP	R(2+1)D	10.7	25.9	35.4	47.3	5.7	19.5	30.7	45.8
	VCP	R(2+1)D	19.9	33.7	42.0	50.5	6.7	21.3	32.7	49.2
	PRP	R(2+1)D	20.3	34.0	41.9	51.7	8.2	25.3	36.2	51.0
	Pace	R(2+1)D	25.6	42.7	51.3	61.3	12.9	31.6	43.2	58.0
	STS	R(2+1)D	38.1	58.9	68.1	77.0	16.4	36.9	50.5	65.4
	VCOP	R3D-18	14.1	30.3	40.0	51.1	7.6	22.9	34.4	48.8
	VCP	R3D-18	18.6	33.6	42.5	53.3	7.6	24.4	36.3	53.6
	PRP	R3D-18	22.8	38.5	46.7	55.2	8.2	25.8	38.5	53.3
	Pace	R3D-18	23.8	38.1	46.4	56.6	9.6	26.9	41.1	56.1
	STS	R3D-18	38.3	59.9	68.9	77.2	18.0	37.2	50.7	64.8
	TransRank	R3D-18	46.5	**63.7**	**72.8**	-	**19.4**	**45.4**	**59.1**	-
Contrastive video SSL methods	MemDPC	R18	20.2	40.4	52.4	64.7	7.7	25.7	40.6	57.5
	MLRep	R3D-18	39.6	57.6	69.2	**78.0**	18.8	39.2	51.0	63.7
	Ours	R3D-18	**48.5**	58.6	65.3	72.1	17.6	35.7	51.4	65.5
		R(2+1)D	47.6	58.8	66.2	74.6	19.0	39.7	54.2	**82.8**

3.4 Loss function

The overall loss function is a combination of the three loss terms introduced above:

$$\mathcal{L} = \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{reorg}}, \tag{7}$$

where the third term only takes effect at the late stage of the training process.

4 Experiments

4.1 Implementation details

Datasets. We conduct experiments on two representative video datasets, namely UCF101[1] (Soomro et al., 2012) and HMDB51[2] (Kuehne et al., 2011). UCF101 consists of 13K videos of 101 action classes. HMDB51 contains 7K videos from 51 action classes. Both UCF101 and HMDB51 have three official train/test splits. We pre-trained on the first train split of UCF101 and used the first train/test split of UCF101 and HMDB-51 for evaluation following (Wang et al., 2020; Qian et al., 2021a). For the ablation study, we use the first train/test split of UCF101.

Network architectures. We experiment on two backbones that are commonly used in previous video SSL methods, namely R3D-18 (Hara et al., 2018) and R(2+1)D-18 (Tran et al., 2018). Given N_p randomly sampled 16-frame clips of resolution 112×112 from a video, the backbone outputs N_p D-dim feature vectors, where $D = 512$, and $N_p = 5$ is the number of types of fixation masks. We sample N_p clips for a video and ensure that all the N_p memory slots for the video can be updated.

Pre-training. The batch size for R3D is 16, and the batch size for R(2+1)D is 14. The two backbones are trained for 300 epochs on the UCF101 training set using SGD with a momentum of 0.9 and weight decay of 10^{-4}. The initial learning rate is 0.1 and is decayed by 5 at epoch 90, 180, and 240. The number of negative samples $N_{\text{neg}} = 1024$. For PE minimization, we set $l = 5$, *i.e.*, the state $\mathbf{c}_i^*$ is from the 5-th level of the backbone network, which shows the best balance between accuracy and efficiency. For reorganization, we use an unsupervised clustering algorithm Faiss (Johnson et al., 2019), and set $R = 3$, $Q^{(r)} = 1500$ for $r = 1, \ldots, R$, and $Q = 1024$ based on ablation study. We train for another 60 epochs after incorporating $\mathcal{L}_{\text{reorg}}$ using SGD with a learning rate of 8^{-4}.

Action recognition. We initialize the backbone using the model parameters obtained in the pre-training stage except for the last linear layer. We consider two settings: i) *linear probe*, where only

[1] https://www.crcv.ucf.edu/data/UCF101.php

[2] https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

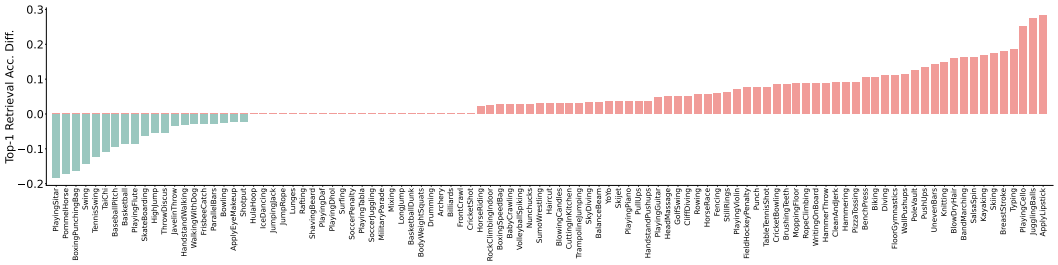

Figure 4: **Top-1 retrieval accuracy comparison before and after reorganization on UCF101.** The reorganization process can improve the Top-1 retrieval performance by 57.4% out of all the classes, leading to comparative performance of 22.8% out of all. See §4 for more discussions.

the last linear layer is trained with cross-entropy loss, and ii) *finetune*, where the entire network is finetuned with cross-entropy loss. For *linear probe*, when training on UCF101, we use SGD optimizer and trained for 200 epochs with an initial learning rate of 0.1, which is further decayed by 10 at epoch 60, 120 and 180. For HMDB51, we use Adam with a learning rate of 0.001 and trained to converge. We use batch size 32 for both R3D and R(2+1)D backbones, and the input resolution is 112×112. For *finetune*, we use an SGD optimizer and trained for 200 epochs with a large initial learning rate of 0.1 following (Duan et al., 2022), which is further decayed by 10 at epoch 60, 120, and 180. The batch size is 32 for both R3D and R(2+1)D backbones. During training, we apply the same data augmentation as in (Han et al., 2020b). For evaluation, we uniformly sample 10 clips from one testing video, perform the center crop, resize them to 112×112, and average the predicted probabilities as the final prediction, following (Wang et al., 2020; Qian et al., 2021a).

4.2 Comparison with the state-of-the-art

To evaluate the representations learned in the self-supervised pre-training stage, we perform *video retrieval* and *action recognition* and compare them with other state-of-the-art methods.

Video retrieval. We query the k-nearest neighbors of the testing set video clips from the training set. The pre-trained backbone model (R3D or R(2+1)D) is directly used as a feature extractor without further finetuning. For each query video, we obtain one 512-d feature vector by extracting and averaging the features of 10 uniformly sampled clips. We use cosine similarity to measure the distance of the features for determining the k-nearest neighbors. A correct retrieval is counted when the k nearest neighbors contain at least one video of the same class with the query video. We report Top-k retrieval accuracy in Table 1, where $k = 1, 5, 10, 20$, and compare it with other video SSL methods pre-trained on the RGB modality of UCF101, including VCOP (Xu et al., 2019), VCP (Luo et al., 2020), PRP (Yao et al., 2020), Pace (Wang et al., 2020), STS (Wang et al., 2021a), and TransRank (Duan et al., 2022). As can be observed, our method achieves the best Top-1 retrieval accuracy on UCF101, and comparable Top-1 performance on HMDB51. Though our method is slightly inferior to other Top-k values, we argue that the Top-1 metric matters the most in security demanded real-world applications such as autonomous driving. Primates including humans are also more good at giving out one promising guess than several candidates (Freedman et al., 2001).

Compared to previous methods, our method captures finer-grained semantics, which are reorganized to yield a better representation structure. To see the change of the learned finer-grained semantics before and after reorganization, we make statistics on the Top-1 retrieval accuracy before and after reorganization on UCF101 regarding the 101 categories. As shown in Fig. 4, the post-learning reorganization improves the Top-1 retrieval accuracy for 57.4% of the 101 action categories. For 22.8% categories, the Top-1 retrieval accuracy remains the same. This shows that the redistributed finer-grained semantics can facilitate similarity based categorization.

We further visualize video retrieval examples that are corrected by reorganization or conversely in Fig. 5. The first two examples are rectified by reorganization. For the `JugglingBalls` query clip, it retrieves a `BoxPunchingBag` clip with similar foreground and background color, *i.e.*, a person in black and red background. For the `PlayingCello` query clip, it retrieves a `PlayingSitar` clip that resembles in both appearance and motion. By representation reorganization, such ambigu-

Figure 5: **Qualitative results for video retrieval before and after reorganization.** The first two examples are improved by reorganization, and the third example is a failure case, where the retrieval result after the reorganization is more similar in appearance while different in semantics. See §4.

Table 2: **Action recognition results on UCF101 and HMDB51.** The models are pre-trained on UCF101 with RGB only. * means learning without back-propagation. See §4.2 for details.

Method	Backbone	Input Size	Frozen	UCF101	HMDB51
CLAPP* (Illing et al., 2021)	VGG-5	-	✓	41.6	-
Object Patch (Hadsell et al., 2006)	AlexNet	-	✓	42.7	15.6
Seq Ver. (Misra et al., 2016)	CaffeNet	-	✓	50.9	19.8
OPN (Lee et al., 2017)	CaffeNet	1×80^2	✓	56.3	22.1
Geometry (Gan et al., 2018)	FlowNet	-	✓	54.1	22.6
Bi-Geometry (Gan et al., 2018)	FlowNet	-	✓	55.1	23.3
Ours	R3D-18	16×112^2	✓	**59.2**	**32.0**
VCOP (Xu et al., 2019)	R3D-18	16×112^2	✗	64.9	29.5
PRP (Yao et al., 2020)	R3D-18	16×112^2	✗	66.5	29.7
STS (Wang et al., 2021a)	R3D-18	16×112^2	✗	70.4	34.9
VCP (Luo et al., 2020)	R3D-18	16×112^2	✗	66.0	31.5
Pace (Wang et al., 2020)	R3D-18	16×112^2	✗	77.1	36.6
MLRep (Qian et al., 2021a)	R3D-18	16×112^2	✗	76.2	41.1
Ours	R3D-18	16×112^2	✗	**76.6**	**43.1**
VCOP (Xu et al., 2019)	R(2+1)D	16×112^2	✗	72.4	30.9
PRP (Yao et al., 2020)	R(2+1)D	16×112^2	✗	72.1	35.0
STS Wang et al. (2021a)	R(2+1)D	16×112^2	✗	77.8	40.7
VCP (Luo et al., 2020)	R(2+1)D	16×112^2	✗	66.3	32.2
Pace (Wang et al., 2020)	R(2+1)D	16×112^2	✗	75.9	35.9
Ours	R(2+1)D	16×112^2	✗	**78.1**	**44.8**

ousness are removed, and clips with the correct action labels are found. However, we also notice that, for clips with complicated backgrounds, such as `PommeiHorse` in the third row of Fig. 5, reorganization tends to make it easier to be mixed up with clips having similar appearance.

Action recognition. We report the Top-1 action recognition accuracy of linear probe (`Frozen` ✓) and finetune (`Frozen` ✗) in Table 2. For a fair comparison, we exclude methods based on much deeper backbones, with larger input resolution, using multi-model data, or pre-trained on much larger video datasets such as Kinetics (Carreira & Zisserman, 2017). For linear probe experiments, our method outperforms all previous video SSL methods pre-trained on UCF101, especially for CLAPP (Illing et al., 2021), a bio-inspired unsupervised representation learning method without back-propagation. This clearly demonstrates that our video SSL method is a competitive practice of exploring cognitive inspirations in deep self-supervised representation learning. For the finetune setting, our method achieves the highest Top-1 accuracy compared with other video SSL methods pre-trained on the RGB modality of UCF101, which shows a good generalization ability of the learned representations.

Table 3: **Ablation study on framework design**. We report Top-1 video retrieval results on UCF101 split 1. Here, N is #training samples and N_p is #fixation locations considered. See §4.3 for details.

No.	Saccade	Pred. Err.	Reorg.	Mem. Size	Top-1
1				N	39.4
2				$N * N_p$	42.5
3	✓			$N * N_p$	44.2
4	✓	✓		$N * N_p$	45.0
5	✓		✓	$N * N_p$	46.5
6	✓	✓	✓	$N * N_p$	**48.5**

Table 4: **Ablation study on reorganization parameters** assessed by Top-1 retrieval accuracy on UCF101 split 1. Here, R is #cluster results, and $Q^{(r)}$ is #clusters in the r-th result. All the baselines are pre-trained for 100 epochs. See §4.3.

No.	R	$Q^{(r)}$	Top-1	No.	R	$Q^{(r)}$	Top-1
1	0		40.8	7	3	500 1000 1500	41.7
2	1	500	41.5	8	3	1000 1000 1000	42.2
3	1	1000	42.1	9	3	1500 1500 1500	**42.3**
4	1	1500	42.0	10	3	1000 1500 2000	41.8
5	1	2000	41.9				
6	1	2500	41.3				

4.3 Ablation study

In this section, we assess the effectiveness of the framework design regarding three components: the semantic-change-aware contrastive learning that utilizes saccade as an indicator (§3.1), the semantic consistency learning by minimizing the prediction error (PE)(§ 3.2), and the reorganization via prototypical contrastive learning (§3.3). We report Top-1 video retrieval accuracy on UCF101 to evaluate the learned video representations without further finetuning.

Overall framework design. As shown in Table 3, although increasing the size of the memory bank benefits the retrieval performance, major improvements are brought by incorporating the three components. Specifically, incorporating saccade for constructing negative pairs achieves an absolute improvement of 1.7 point, and further including PE minimization or post-learning reorganization leads to an absolute improvement of 0.8 and 2.4, respectively. The full model is 4.3 point higher than the baseline with a memory bank of the same size. The results clearly demonstrate the effectiveness of each framework design.

Reorganization parameters. To better reorganize the learned finer-grained semantics, we experiment on two key parameters introduced in §3.3, namely the number of cluster results R and the number of clusters in each result $Q^{(r)}$. All the baselines are trained on UCF101 for 100 epochs using SGD. The initial learning rate is 0.1 and is decayed by 5 at epoch $30, 60$ and 80. The reorganization loss is incorporated at epoch 61, and the prototypes are updated every 5 epochs. As shown in Table 4, reorganization can consistently improve the Top-1 retrieval accuracy for a wide range of parameter settings. In our full experiments where the models are trained for 300 epochs, we set $R=3$ and $Q^{(r)}=1500$ for $r=1, \cdots, R$.

5 Conclusion

In this work, we propose a video SSL method by taking inspiration from cognitive science and neuroscience on human visual perception. Instead of designing pretext tasks based on the inherent properties of videos, we explore the presence of saccades as an indicator of semantic change in a contrastive learning framework to mimic the role of self-awareness in human perception. To achieve semantic consistency in the absence of a saccade, we minimize the prediction error when using the state of a time point to predict that of another time point during a fixation. Finally, we strengthen the associations among similar representations through a post-learning reorganization process. Compared to previous contrastive learning based video SSL methods, our method learns more powerful representations by first making finer-grained discrimination for semantics in a video instance with temporal consistency and then associating similar semantics across different video instances through a reorganization process. The proposed bio-inspired video SSL method shows superior Top-1 video retrieval accuracy on UCF101 and outperforms other methods on the action recognition tasks.

REFERENCES

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 33:25–37, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Nadine Behrmann, Jurgen Gall, and Mehdi Noroozi. Unsupervised video representation learning by bidirectional feature prediction. In *WACV*, pp. 1670–1679, 2021.

Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *CVPR*, pp. 9922–9931, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 6299–6308, 2017.

Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *AAAI*, volume 1, 2021.

Hanneke EM Den Ouden, Peter Kok, and Floris P De Lange. How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, 3:548, 2012.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.

Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, and Jue Wang. Motion-aware self-supervised video representation learning via foreground-background merging. In *CVPR*, 2022.

Haodong Duan, Nanxuan Zhao, Kai Chen, and Dahua Lin. Transrank: Self-supervised video representation learning via ranking-based transformation recognition. In *CVPR*, pp. 3000–3010, 2022.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *CVPR*, pp. 1801–1810, 2019.

Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, pp. 3636–3645, 2017.

David J Freedman, Maximilian Riesenhuber, Tomaso Poggio, and Earl K Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, 2001.

Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, pp. 5589–5597, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pp. 1735–1742, 2006.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCVW*, pp. 0–0, 2019.

Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, pp. 312–329, 2020a.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *NeurIPS*, 2020b.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pp. 6546–6555, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. In *ICCV*, 2021.

Jarmo Hurri and Aapo Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691, 2003.

Bernd Illing, Jean Ventura, Guillaume Bellec, and Wulfram Gerstner. Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *NeurIPS*, 34, 2021.

Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *ECCV*, pp. 425–442, 2020.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Trans. on Big Data*, 7(3):535–547, 2019.

Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, volume 33, pp. 8545–8552, 2019.

Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. Cycle-contrast for self-supervised video representation learning. *NeurIPS*, 33:8089–8100, 2020.

Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *ICCVW*, pp. 3195–3204, 2021.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *CVPR*, pp. 2556–2563, 2011.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *CVPR*, pp. 667–676, 2017.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2020.

Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Motion-focused contrastive learning of video representations. In *ICCV*, pp. 2105–2114, 2021.

Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *ICCV*, pp. 8239–8249, 2021.

Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *AAAI*, volume 34, pp. 11701–11708, 2020.

Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *CVPR*, pp. 2203–2212, 2017.

David Melcher and Carol L Colby. Trans-saccadic perception. *Trends in Cognitive Sciences*, 12 (12):466–473, 2008.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pp. 527–544, 2016.

Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *ICML*, pp. 737–744, 2009.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, pp. 11205–11214, 2021.

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, pp. 2701–2710, 2017.

Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *ICCV*, pp. 9577–9587, 2021.

Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. In *ICCV*, pp. 7990–8001, 2021a.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, pp. 6964–6974, 2021b.

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.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, pp. 843–852, 2015.

Chen Sun, Arsha Nagrani, Yonglong Tian, and Cordelia Schmid. Composable augmentation encoding for video representation learning. In *ICCV*, pp. 8834–8844, 2021.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pp. 6450–6459, 2018.

Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, pp. 98–106, 2016.

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *ECCV*, pp. 391–408, 2018.

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 *CVPR*, pp. 4006–4015, 2019a.

Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, pp. 504–521, 2020.

Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Wei Liu, and Yun-Hui Liu. Self-supervised video representation learning by uncovering spatio-temporal statistics. *IEEE TPAMI*, 2021a.

Jinpeng Wang, Yuting Gao, Ke Li, Xinyang Jiang, Xiaowei Guo, Rongrong Ji, and Xing Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *AAAI*, 2021b.

Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *CVPR*, pp. 11804–11813, 2021c.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pp. 2794–2802, 2015.

Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, pp. 2566–2576, 2019b.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *CVPR*, pp. 8052–8060, 2018.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pp. 3733–3742, 2018.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, pp. 16684–16693, 2021.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, pp. 10334–10343, 2019.

Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*, volume 35, pp. 10656–10664, 2021.

Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, pp. 6548–6557, 2020.