PreViTS: Contrastive Pretraining with Video Tracking Supervision

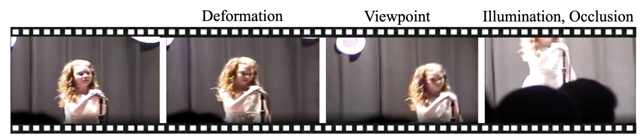
Brian Chen[1] Ramprasaath R. Selvaraju[2] Shih-Fu Chang[1] Juan Carlos Niebles [3] Nikhil Naik [3]

[1]Columbia University [2]Artera AI [3]Salesforce Research,

{bc2754,sc250}@columbia.edu ram@artera.ai {jniebles,nnaik}@salesforce.com

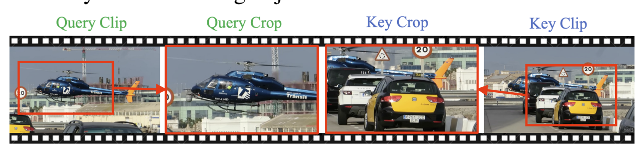
Abstract

Videos are a rich source for self-supervised learning (SSL) of visual representations due to the presence of natural temporal transformations of objects. However, current methods typically randomly sample video clips for learning, which results in an imperfect supervisory signal. In this work, we propose PreViTS, an SSL framework that utilizes an unsupervised tracking signal for selecting clips containing the same object, which helps better utilize temporal transformations of objects. PreViTS further uses the tracking signal to spatially constrain the frame regions to learn from and trains the model to locate meaningful objects by providing supervision on Grad-CAM attention maps. To evaluate our approach, we train a momentum contrastive (MoCo) encoder on VGG-Sound and Kinetics-400 datasets with PreViTS. Training with PreViTS outperforms representations learnt by contrastive strategy alone on video downstream tasks, obtaining state-of-the-art performance on action classification. PreViTS helps learn feature representations that are more robust to changes in background and context, as seen by experiments on datasets with background changes. Our experiment also demonstrates various visual transformation invariance captured by our model. Learning from large-scale videos with PreViTS could lead to more accurate and robust visual feature representations.

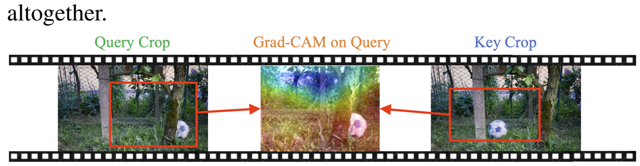


(a) Temporal transformations in videos provide a natural source of data augmentation, making them attractive for self-supervised learning (SSL).

(b) Randomly selected query and key clips in contrastive video SSL may lead to missing objects.



(c) Query and key clips may also contain different visual concepts altogether.



(d) Since many videos contain a fixed background, SSL models can cheat by focusing on the background.

Figure 1: Current methods for contrastive video self-supervised learning receive an imperfect supervisory signal and can rely on background correlations when learning representations. We propose a new approach by video tracking and Grad-CAM supervision to tackle these problems.

1. Introduction

Self-supervised learning (SSL) of visual representations [58, 63, 35, 50, 20, 9, 10, 11] has become a competitive alternative to supervised learning, without requiring manually annotated labels. A key component of SSL from images is contrastive learning, a learning objective that pulls different data augmentations from the same instances (known as query and key) to be closer to each other and pushes data augmentations from different instances away. However, not all of the commonly used augmentations in images reflect the visual variability that we see in the real world.

In contrast, videos provide a natural source of data augmentation, with objects undergoing deformations and occlusions, along with changes in viewpoints and illumination as shown in Figure 1a. As a result, recent work has tackled SSL from videos to seek more natural augmentations and meaningful semantics [38, 36, 32, 56, 29, 5, 57, 27, 45, 41, 25]. A common approach [17, 8] is to randomly sample nearby clips in videos as query and key as a natural way of data augmentation that represents the same instance since frames that are close in time are likely to share similar content. However, this sampling strategy for augmentation suffers from a few problems, as shown in Figure 1b and 1c. First, when sampling instances from a longer span of the video, the content might change substantially, resulting in samples containing totally different semantic concepts. This sampling strategy results in an imperfect supervisory signal that does not encourage semantic understanding. Second, when sampling clips from the same video, the background in the two clips are often quite similar, which allows the model to cheat by looking at the background for minimizing contrastive loss [52] as shown in Figure 1d. This sampling strategy leads to models learning spurious background correlations and context, which could make them less transferable and potentially biased [12].

To alleviate these problems, we propose Pretraining with Video Tracking Supervision (PreViTS). PreViTS consists of an intelligent method to select query and key clips, which utilizes unsupervised tracking for videos. Using this freely available form of supervision, we design a temporal constraint for selecting clips that ensures that the query and the key contain the same object. In addition, using tracking information on the spatial extent of the object, we design spatial constraints to mask the background. Taken together, these spatial-temporal constraints result in better supervisory signals for contrastive learning from videos. After selecting more informative query and key clips, we train the model to learn to localize specific regions in query and key that represent the same concepts using a Grad-CAM [47]-based attention loss. We pretrained a momentum contrastive encoder (MoCo) [20] with PreViTS on Image and Video-based SSL backbones using VGG-Sound and Kinetics-400 datasets. Evaluation on video downstream tasks, including action recognition, video retrieval shows that PreViTS-trained models learn more accurate visual representations. In particular, we obtain state-of-the-art performance on video action classification. Due to its ability to localize objects, PreViTS-trained models can perform unsupervised tracking across arbitrary lengths of videos, as shown by our experiments on the DAVIS challenge [40]. Additional experiments on image and video datasets with background changes show that models trained with PreViTS are less dependent on background correlations and are more robust to background changes in visual classification. We also showed the various invariances (occlusion, viewpoint) captured by our model.

In sum, our work shows that existing methods for contrastive SSL from videos do not efficiently use temporal transformations of objects. By designing a better clip sampling strategy and a loss that encourages object localization, we are able to learn more accurate visual representations from the video that are robust to background changes.

2. Related Work

Self-supervised representation learning (SSL). Contrastive SSL approaches learn image representations [2, 21, 50, 64, 9, 31, 3] by forming positive and negative pairs, and maximizing the similarity of positive pairs as compared to negative pairs. Positive pairs are generated from a single image instance through artificial data augmentations such as random cropping, resizing, color distortion, and Gaussian blur [9]. Going beyond learning representations from images, different frames of videos provide natural viewpoint changes and temporal information which can help learn better representations in a self-supervised manner [1, 55, 38, 56, 51, 42, 44, 23, 43]. Saliently, contrastive learning-based methods [17, 41, 25, 62, 15] that sample positive pairs from the same video have shown that view-point invariant representations can be learnt from videos. Unlike previous methods [55, 41] that sample positive pairs from unsupervised proposals with bounding boxes, we introduce an approach for sampling pairs based on spatial and temporal constraints obtained using unsupervised saliency maps, coupled with Grad-CAM supervision [47] to learn better grounded representations.

Grounded Representation Learning. Our work is also related to recent work on learning better grounded representations. Henaff *et al.* [22] introduced DetCon, a self-supervised objective which tasks representations with identifying object-level features across different image augmentations. Mo *et al.* [33] introduced a technique to mix backgrounds of different images during contrastive pretraining and showed that it leads to models learning reduced contextual and background biases. Xie *et al.* [60] propose an object-level pretraining approach for learning from complex scenes. CAST [48] learns visually grounded representations through saliency supervision. FAME [14] extracts moving foreground by frame difference and color statistics to alleviate background bias.

3. Method

We propose Pretraining with Video Tracking Supervision (PreViTS) to learn visual representations from videos by utilizing unsupervised object tracking. First, we will review the standard contrastive based video representation learning framework and then discuss our approach.

3.1. Background

When performing contrastive learning on videos, the positive pairs are clips from the same video selected from different times, while the negative pairs are formed with clips taken from other videos. In this work, we build our approach on top of the Momentum Contrast (MoCo) [20] model, which uses the InfoNCE [35] objective and stores

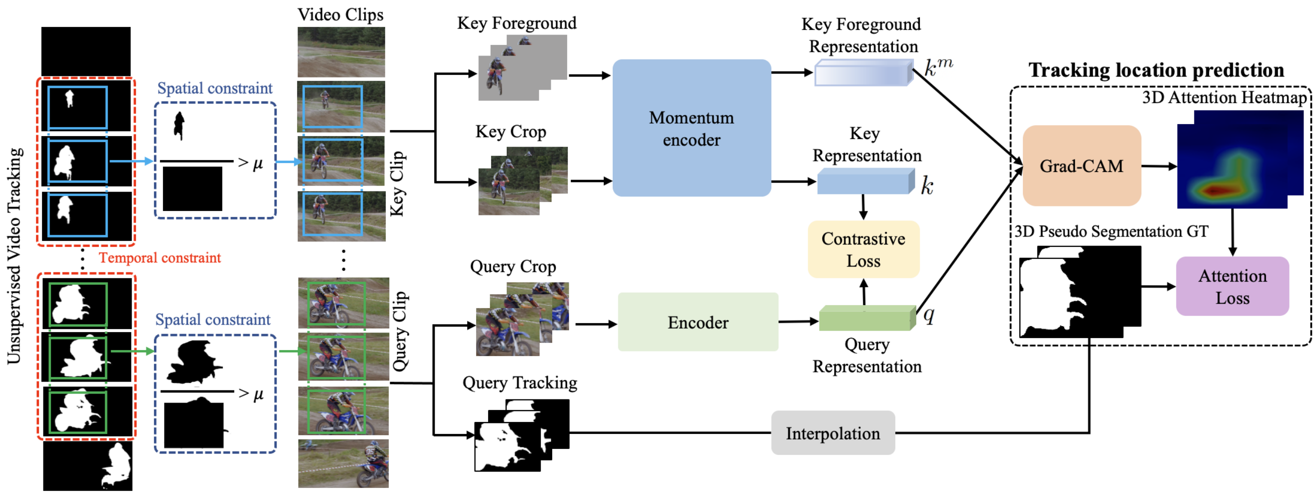


Figure 2: **Pretraining with Video Tracking Supervision (PreViTS):** Given an input video, we perform unsupervised tracking and apply temporal constraints to extract continuous frames that contain the tracked object region. We then apply IoU based spatial constraints to sample query and key video clips along with their masks. The encoder representations for the query and key are aligned through a contrastive loss. We then mask the key and use Grad-CAM to localize the regions in the query that maximize the (key foreground, query) similarity. We then supervise Grad-CAM with the tracked query mask using a cosine distance loss to encourage models to rely on appropriate salient object regions during contrastive pretraining.

the negative samples in a dynamic memory bank with a moving average encoder. Formally, given a video V, we learn feature representations for query q and key k sampled from the same video. The goal is to pull the feature distance of the positive pairs q and k to be closer and push the features of query q away from a negative set of features from other videos $N = \{n_1, n_2, ..., n_m\}$. The MoCo loss is:

$$\mathcal{L}_{\text{MoCo}} = -\log \frac{\exp{(q \cdot k)/\tau)}}{\sum_{n \in \{N,k\}} \exp{(q \cdot n)/\tau)}}, \quad (1)$$

where τ is the temperature constant.

In the video model, in addition to the MoCo loss, we also use the **relative speed prediction task** which has been found to be beneficial to understand the relative speed between the video segments proposed in **RSPNet** [8]. We sample three video segments: two segments with the same speed and another with a different speed. The goal is to pull the feature distance for segments with the same speed closer together while pushing the features for the segment with different speed away. A triplet loss [46] is applied:

$$\mathcal{L}_{Speed} = \max(0, \gamma - (pair^{+} - pair^{-})), \quad (2)$$

where the distance of positive pairs $pair^{+}$ should be larger than the negative pairs $pair^{-}$ by a margin $\gamma > 0$.

3.2. Unsupervised tracking in videos

In order to select query and key clips from the same video that contain the same visual concepts, we propose to use unsupervised object tracking to guide clip selection. To acquire unsupervised tracking information from the video we first use Deep-USPS [34], an unsupervised saliency prediction algorithm, to obtain a saliency map for the initial frame in the video. We use this saliency map as the target object for tracking and apply SORT [4], a tracking algorithm which checks the IoU constraint across continuous frame masks to track the target object through the video. Formally, given an input video V with height h, width w and temporal length t, we acquire the video object segmentation map $M \in \{0,1\}^{h \times w \times t}$, where $M_{ijk} = 1$ indicates pixel (i, j, k) is salient, and area of salient region in time t is $A_M^t = \sum_{i,j} M_{i,j}$. The saliency map is a binary mask. Since a large majority of the web videos (and as a result, videos in vision datasets) are centered on a single object, we only utilize one (the largest) salient region in the video for tracking and do not consider multiple objects in this work.

Spatial-temporal cropping based on video tracking: Once we obtain the tracking tube for the video, we constrain our random sampling to video segments covered by the tracking tube as shown in left half of Figure 2, where $A_M^t \neq 0$. This ensures that our sampled query and key clips contain meaningful instances of the same object in the video. In addition, we set a spatial constraint (Figure 2): the random crop for the query or key should have at least $\mu \in [0, 1)$ IoU with the tracking mask. This spatial constraint tries to ensure that the query and key contain the same object for contrastive pretraining. We acquire two 3D masks for the video segment M_q and M_k, which represent the mask of the *query* and *key* containing salient regions.

3.3. Pretraining with Video Tracking Supervision (PreViTS)

PreViTS aims to encourage the model to learn to localize specific regions within the query and key that represent the same concept. We first determine the regions that the network relies on when matching the object regions in the key, x^k with that of the query, x^q. To obtain the object regions in key, we mask the key with the video segmentation mask, M_k, as a filter to get the key foreground, $x^{k_m} = x^k * M_k$. To understand the importance placed by the network on specific crop regions when contrastively matching their representations, we compute Grad-CAM [47] in a contrastively-trained fashion. We do this by first forward propagating the key foreground, x^{k_m}, and the query, x^q, through the respective encoders to get k^m and q. To get the regions that would help maximizing their similarity, we take their dot-product and compute the gradients *wrt* the last convolution layer activations of the query encoder, f_q, as follows:

$$\alpha_q = \overbrace{\sum_{i,j}}^{\text{global pooling}} \underbrace{\frac{\partial q \cdot k^m}{\partial A^{f_q}_{conv5}}}_{\text{gradients via backprop}} \tag{3}$$

where the α_q represents the last convolutional layer neurons' importance for maximizing the similarity of the query and the key foreground representations. Through a weighted combination of α_q with the last convolutional layer activations $A^{f_q}_{conv5}$ and clipping them at zero, we can get Grad-CAM maps, G_q.

$$\mathrm{G}_q = ReLU \underbrace{\left(\sum_n \alpha_{\mathrm{q}} A^{f_{\mathrm{q}}}_{conv5} \right)}_{\text{linear combination}}. \tag{4}$$

Higher values in G_q represents the regions the network relies on when mapping query to key foreground.

We would ideally want the network to only rely on the tracked object regions in the query that are highlighted in the key foreground. Therefore, we apply a cosine-distance based attention loss to encourage the Grad-CAM heatmap G_q to be close to tracked object mask in the query segment M_q. This enforces the model to learn similar representations for the object irrespective of the viewpoint and transformation changes that might be present in the clips when the frames are temporally far away. We interpolate M_q to the same spatial and temporal dimension as G_q to acquire the pseudo segmentation ground-truth, $\hat{M}_q$ as the supervision for the Grad-CAM heatmap. The Attention loss is defined as:

$$\mathcal{L}_{\text{att}} = 1 - \frac{G_q \cdot \hat{M}_{\text{q}}}{\|G_q\| \, \|\hat{M}_{\text{q}}\|}. \tag{5}$$

Our full model is trained to minimize the sum of the losses described above.

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MoCo}} + \mathcal{L}_{\text{Speed}} + \lambda \mathcal{L}_{\text{Att}}. \tag{6}$$

4. Experiments

We aim to show that training video self-supervised models with PreViTS leads to better representations that obtain improved transfer learning performance with reduced dependence on background signal and context. We validate this by pretraining representations on two datasets and transferring them to various video and tracking tasks.

4.1. Implementation details

We pretrain our models on two datasets independently, both consist of 10 second-long videos at 25 FPS: (1) The **VGG-Sound** [7] dataset contains 200k videos collected from YouTube. VGG-Sound was collected with the objective of creating an audio-visual dataset with diverse sounds and contains 300 classes as defined by audio labels. Unlike previous video SSL methods that test on video downstream tasks, we also learn object concepts from videos for image understanding. So we chose VGG-Sound, which contains a wider variety of object classes and higher object-centricity as compared to action classification datasets common in the video understanding literature. Also, a large majority of VGG-Sound videos only contain a single foreground object, as we found by using supervised segmentation, which is consistent with our single object assumption in the learning phase. (2) The **Kinetics-400** dataset [6] is a widely-used dataset, which enables us to compare PreViTS's performance to prior methods. It consists of around 240k training videos with 400 human action classes. We will release the code for replicating our work. More details and image recognition experiments can be found in the supplement.

Method	**Dataset**	**UCF-101**
RSPNet	VGG Sound	86.4
+ Tracking Constrained Sampling	VGG Sound	87.5 $_{+1.1}$
+ PreViTS	VGG Sound	88.9 $_{+2.5}$
RSPNet	K400	87.6
+ Tracking Constrained Sampling	K400	89.1 $_{+1.5}$
+ PreViTS	K400	91.8 $_{+4.2}$

Table 1: **Video Action Classification:** PreViTS obtains significant performance gains on the commonly-evaluated downstream task of UCF-101 action recognition. Tracking Constrained Sampling refers to our unsupervised tracking-based spatial-temporal sampling strategy.

4.2. Video tasks

Action recognition: To evaluate the performance of PreViTS-trained models on video classification tasks, we perform action recognition on the UCF-101 dataset [49]. Following Xu *et al.* [61], in all experiments, we finetune

Method	Input size	Params	Backbone	UCF-101
RSPNet [8]	112 × 112	33.4M	C3D	76.7
CACL [18]	112 × 112	33.4M	C3D	77.5
PreViTS	112 × 112	33.4M	C3D	**78.7**
Pace [54]	112 × 112	14.4M	R(2+1)D	77.1
STS [53]	112 × 112	14.4M	R(2+1)D	77.8
VideoMoCo [37]	112 × 112	14.4M	R(2+1)D	78.7
RSPNet [8]	112 × 112	14.4M	R(2+1)D	81.1
PreViTS	112 × 112	14.4M	R(2+1)D	**81.9**
SpeedNet [3]	224 × 224	9.6M	S3D-g	81.1
CoCLR [19]	224 × 224	9.6M	S3D-g	87.9
STS [53]	224 × 224	9.6M	S3D-g	89.0
RSPNet [8]	224 × 224	9.6M	S3D-g	89.6
ASCNet [24]	224 × 224	9.6M	S3D-g	90.8
PreViTS	224 × 224	9.6M	S3D-g	**91.8**

Table 2: **Comparison to prior work on UCF-101 performance:** Our best-model trained with PreViTS outperforms all existing methods for video self-supervised learning on UCF-101 downstream performance, when using comparable training resources.

Method	Top-k				
	$k=1$	$k=5$	$k=10$	$k=20$	$k=50$
Pace [54]	31.9	49.7	59.2	68.9	80.2
RSPNet [8]	36.0	56.7	66.5	76.3	87.7
STS [53]	39.1	59.2	68.8	77.6	86.4
CACL [18]	43.2	61.1	69.9	78.2	88.2
TCLR [13]	48.6	67.6	75.5	82.5	-
PreViTS	**53.4**	**69.4**	**77.8**	**85.5**	**93.0**

Table 3: **Video retrieval** results on UCF101. Our model outperforms other baselines using the same architecture C3D backbone.

our pretrained model on labeled videos with 50 epochs using a learning rate of 0.05. We drop the projection head and replace it with a randomly initialized fully-connected layer. We report top-1 accuracy on the UCF-101 dataset when pretraining with PreViTS on VGG-Sound and Kinetics-400 datasets in Table 1. Training with PreViTS obtains a substantial improvement over RSPNet on both pretraining datasets. Notably, the model pretrained on Kinetics-400 had better performance with RSPNet and a larger absolute improvement with RSPNet + PreViTS (4.2% versus 2.5%), over VGG-Sound. We speculate that since human actions are better represented in Kinetics-400, the representation learnt using these videos transfers better to UCF-101, and also benefits more from training with PreViTS. Finally, we compare the performance of RSPNet + PreViTS pretrained with Kinetics-400 with other state-of-the-art video SSL methods [8] in Table 2. With the same architecture, computational budget, epoch, batch size, and pretraining data for a fair comparison, our approach outperforms prior work and obtains state-of-the-art performance.

Video retrieval: We also evaluate our video retrieval task

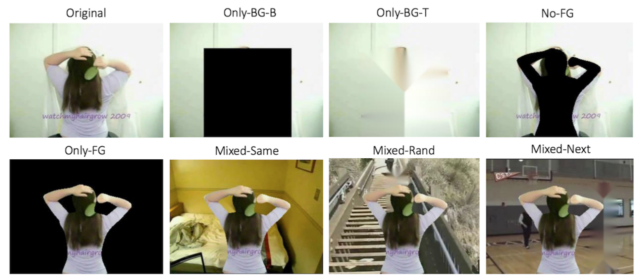


Figure 3: **Video Background Challenge:** We evaluate PreViTS by introducing a Video Backgrounds Challenge to evaluate background-robustness of video models. FG = foreground, BG = background. Foreground-background combinations include: Only-BG-B (FG: Black, BG: Unmodified), Only-BG-T (FG: Tiled background, BG: Unmodified), Mixed-Same (FG: Unmodified, BG: Random BG of the same class), Mixed-Rand (FG: Unmodified, BG: Random BG of a random class), and Mixed-Next (FG: Unmodified, BG: Random BG of the next class.)

on the UCF-101 dataset. Given a video as a query, we search the most relevant video by cosine distance using the nearest neighbor search. Following [8], we evaluate our method on the split 1 of UCF101 dataset and apply the top-k accuracies (k=1, 5, 10, 20, 50) as evaluation metrics. As shown in Table 3, our model outperforms the other baselines by a large margin, showing the effectiveness of the proposed training process.

4.3. Backgrounds challenge

We expect feature representations obtained using PreViTS to be less dependent on object backgrounds and context. To quantify this, we utilize the "backgrounds challenge" [59] on both image and video classification tasks as shown in Table 4.

Backgrounds Challenge. First, we evaluate our model on the original Backgrounds Challenge [59], which was designed to test a model's robustness to various background changes. It contains 9 ImageNet classes with 450 images for each class. We evaluate our model along with the baseline model pretrained on VGG-Sound and train a linear layer with ImageNet-1K. Results show that pretraining with PreViTS achieves significant improvement on all tasks defined in the Backgrounds Challenge. Examples of different settings can be found in Figure 3. In the Only-FG setting, where the background is set to black, PreViTS obtains an absolute improvement of 12.1%, showing that it is less dependent on background information. When backgrounds are swapped (Mixed-Same, Mixed-Rand, Mixed-Next), PreViTS obtains an absolute improvement of $3.6-4.2\%$, indicating that representations learnt with PreViTS reduce the reliance on background correlations. There is a

	Backgrounds Challenge [59]							
VGG Sound	Original	Mixed-Same	Mixed-Rand	Mixed-Next	Only-FG	No-FG	Only-BG-B	Only-BG-T
MoCo	77.9	53.3	37.8	33.8	40.9	24.6	9.7	13.5
PreViTS	81.0$_{+3.1}$	56.9$_{+3.6}$	42.0$_{+4.2}$	38.0$_{+4.2}$	53.0$_{+12.1}$	28.0$_{+3.4}$	8.8$_{-0.9}$	13.0$_{-0.5}$
K400	**Video Backgrounds Challenge**							
RSPNet	70.7	40.7	30.3	29.5	20.9	49.1	35.2	28.6
PreViTS	74.0$_{+3.3}$	48.0$_{+7.3}$	35.9$_{+5.6}$	32.7$_{+3.2}$	27.8$_{+6.9}$	51.9$_{+2.8}$	33.7$_{-1.5}$	28.3$_{-0.3}$

Table 4: **Robustness to background changes**. On image and video Backgrounds Challenge datasets, PreViTS outperforms baselines where the foreground was included (columns 1-5), especially the Only-FG setting. Also, PreViTS-trained models are less accurate when foreground information is entirely eliminated (columns 7, 8), showing their reduced reliance on background information.

Method	**Occlusion**		**Viewpoint**		**Illumination Dir.**		**Illumination Color**		**Instance**		**Instance+Viewpoint**	
	Top-10	Top-25	Top-10	Top-25	Top-10	Top-25	Top-10	Top-25	Top-10	Top-25	Top-10	Top-25
MOCO	83.25	76.45	84.83	75.31	85.09	74.74	99.42	95.88	48.99	43.55	51.23	46.83
Region Tracker [41]	83.26	76.52	84.97	76.18	88.30	79.34	99.77	97.70	48.81	44.38	53.31	49.04
PreViTS	**85.11**	**78.84**	**89.35**	**81.28**	**91.66**	**83.94**	**99.92**	**98.89**	**55.45**	**49.09**	**56.97**	**51.70**

Table 5: **Invariances of Video representations:** The representation learned by PreViTS is more invariant to various transformations as compared to baseline MoCo, as shown by the top-k Representation Invariance Score (RIS) [41]. The large improvement in viewpoint invariance is likely due to our strategy of sampling tracked objects with different viewpoints. The large improvement in instance invariance shows that PreViTS is better at learning object concepts instead of low-level pixel similarities. Improved invariance is useful for object recognition tasks. See Section 4.4 for details of RIS.

slight increase in performance in the No-FG setting, likely due to the model learning contour information from videos. However, in settings where no information from the foreground is provided (Only-BG-B, and Only-BG-T), PreViTS obtains lower accuracy than baseline, which reinforces that it is less dependent on the background signal.

Video Backgrounds Challenge (JHMDB). Taking inspiration from the Backgrounds Challenge dataset, we construct a new Video Backgrounds Challenge to test background-robustness on videos. We use the JHMDB dataset [26]—consisting of 21 HMDB [28] action recognition classes with 50 videos per class—for which the ground truth foreground mask is available. We follow Xiao et al. [59] to construct 8 foreground-background combinations (Figure 3) for JHMBD. We evaluate performance using a model trained on Kinetics-400 and finetuned on UCF-101 and JHMDB. Models trained with PreViTS outperform the baseline model (RSPNet) in all settings. Similar to the trends on Backgrounds Challenge, PreViTS obtains significant improvement in settings where the background is set to black or is replaced by background from another video. In settings where the foreground is removed, we find the accuracy drop to be higher for PreViTS compared to baseline (22.1 vs. 21.6). Video representation learning models have been shown to suffer from over-reliance on background information, called representation bias [30] or scene bias [12]. Training with PreViTS can help mitigate this bias.

4.4. Invariances captured by PreViTS.

We expect representations learnt by PreViTS to have better invariance to various transformations (occlusion, viewpoint, illumination, instance), due to more effective use of object instance information during contrastive learning. Following [41], we measured the representation's invariances when predicting classes using the top-k Representation Invariance Score (RIS). We selected top-10/25 neurons from encoder with similar activation behavior between transformations and computed its mean score. PreViTS is significantly more invariant to transformations than other baselines (Table 5).

4.5. Video tracking evaluation

To demonstrate grounding and tracking ability, we evaluate PreViTS on single object video tracking [40] in Grad-CAM attention fashion. In the original video tracking task, the input is the first frame of the video along with the foreground segmentation mask. The goal is to predict the pixel-level mask of the foreground in the later video frames. In our setting (Figure 2), we feed the first frame and its segmentation to acquire the key foreground. Then, we feed the later frames as queries and compute the Grad-CAM at-

Region Similarity $\mathcal{J}$	Mean $\mathcal{M}\uparrow$	Recall $\mathcal{O}\uparrow$	Decay $\mathcal{D}\downarrow$
MoCo	0.315	0.638	0.025
PreViTS	0.544	0.769	-0.014

Table 6: **Unsupervised Tracking on DAVIS 2016:** We show that through our grounding supervision, we are able to better track objects across videos of arbitrary lengths given just the first frame and its associated segmentation map.

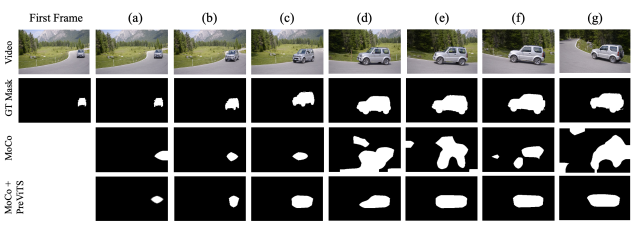


Figure 4: **Unsupervised Object tracking**. Using Grad-CAM attention and the query-key framework, PreViTS-trained model can be used to track objects across the video given the first frame and corresponding segmentation map of the object to track. PreViTS is able to localize objects under viewpoint changes, while the baseline model is unable to do so.

tention heatmap to localize the corresponding region in the later frames. Since the attention heatmap resolution is 7×7, we cannot perform pixel-level prediction. Our evaluation metrics follow [40]. We compute: Region similarity ($\mathcal{J}$), which represents the IoU between the predicted foreground mask and GT foreground mask; Mean ($\mathcal{M}$) is the average value of $\mathcal{J}$; Recall ($\mathcal{O}$) evaluates the fraction of sequences scoring higher than a threshold; Decay ($\mathcal{D}$) evaluates the averaged performance drop over time, e.g., $\mathcal{J}_{t=4} - \mathcal{J}_{t=1}$. PreViTS outperforms the baseline MoCo by a significant margin (Table 6), which demonstrates our model's ability to localize objects in dynamic videos. Figure 4 shows how PreViTS is able to localize objects while the baseline fails when the object appears in a novel viewpoint (Figure 4(d)).

4.6. Ablations and Analysis

We conduct an ablation study on the effect of our design decisions. We evaluate UCF-101 trained on K400 for 50 epochs following [39]. We also tested the image dataset PASCAL VOC object detection [16] trained on the VGG-Sound for 200 epochs. More details of the image model can be found in the supplement.

Temporal distance constraint between positive pairs: We investigate the effect of different temporal sampling strategies in Table 7a. We define δ to be the temporal distance between the query and key segment. $\delta = 0$ uses the same sample segment for query and key with image augmentation. **Constant** δ samples query and key in a fixed length of 1.7 sec, which ends up as an easier task for the model and does not generalize to the downstream task. **Varying** δ does not constrain the distance between the clips, which refers to random sampling query and key without additional temporal distance constraint. We find this setting to perform the best as it enables the network to localize regions across the clips irrespective of their temporal distance.

Effect of area threshold μ**:** We apply spatial constraint when sampling our positive pairs where the crop covers at least μ IoU of the tracking object area. Here, we investigate the different values of μ in the range 0 to 0.9. Results in Table 7b demonstrate that adding spatial constraint helps the model focus on meaningful objects in the video. We also find the performance result is not sensitive to value of μ, demonstrating consistent improvement of our method.

Effect of loss weight λ**:** We test different loss weights λ

Temporal Sampling	Varying δ	Constant δ	$\delta=0$
VOC07	73.0	$72.4_{-0.6}$	$67.5_{-5.5}$
UCF-101	84.5	$83.7_{-1.8}$	$84.3_{-0.2}$

(a) Effect of different temporal sampling strategy.

Spatial area threshold	$\mu=0.0$	$\mu=0.2$	$\mu=0.3$	$\mu=0.4$
VOC07	$71.5_{-1.5}$	$72.1_{-0.9}$	73.0	$72.8_{-0.2}$
UCF-101	$83.7_{-3.7}$	$85.1_{+0.6}$	84.5	$84.2_{-0.3}$

(b) Effect of Area threshold μ (Fixing $\mu=0.3$)

Loss weighing factor	$\lambda=0.0$	$\lambda=2.0$	$\lambda=3.0$	$\lambda=4.0$
VOC07	$70.3_{-2.7}$	$72.4_{-0.6}$	73.0	$72.6_{-0.4}$
UCF-101	$80.8_{-3.7}$	$83.4_{-2.1}$	84.5	$84.1_{-0.6}$

(c) Effect of loss weighing factor λ (Fixing $\lambda=3.0$)

Tracking	No Tracking	Unsup. Box	Unsup. Mask	Sup. Seg
VOC07	$68.3_{-4.7}$	$71.9_{-1.1}$	73.0	$75.0_{+2.0}$
UCF-101	$79.0_{-5.5}$	$83.0_{-1.5}$	84.5	$86.1_{+1.6}$

(d) Effect of different tracking supervision

Table 7: **Ablations for PreViTS training:** We isolate the effects of our training components. We find that **(a)** starting with a shorter temporal distance between query-key clips and relaxing the constraint as training progresses improves performance. **(b)** adding some amount of spatial constraints based on IoU with tracking mask ensures that different clips contain common salient regions and this improves performance. **(c)** increasing weights on attention loss increases the downstream performance up to a certain point, **(d)** replacing unsupervised video tracking supervision with a noisy bounding box tracking tube achieved a significant gain over the baseline. Apply supervised tracking improves downstream performance slightly.

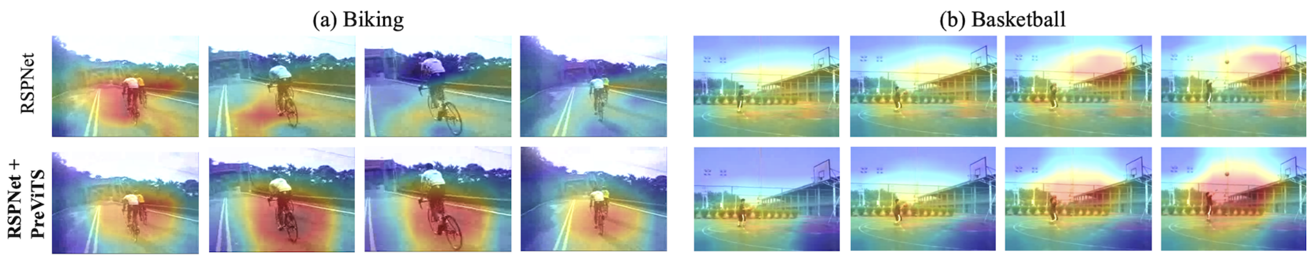


Figure 5: **Visual Grounding for Action Classification.** PreViTS provides better visual grounding as shown by Grad-CAM attention maps of pretrained models finetuned on UCF-101. In (a), our model focuses on the human and bike while the baseline model attends to seemingly irrelevant regions, including the road in the background. In (b), our model attends to the man and the ball in the air in addition to the basketball court while the baseline model focuses mostly on the court.

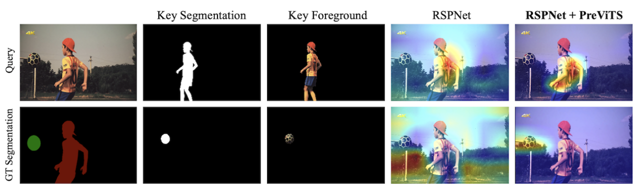


Figure 6: **Discriminative localization of objects.** When provided query with two different segmentation corresponding to different foreground objects and key foregrounds, PreViTS-trained model is able to localize the object accurately, capturing class-specific semantic discrimination between objects.

to balance between the two losses. Results from Table 7b show that non-zero values of λ outperform λ = 0.0, indicating that attention loss is important in PreViTS. Higher λ improves performance up to a point—performance improves with λ= 2.0, 3.0, and slightly degrades with λ= 4.0. We find $\lambda = 3.0$ to be optimal.

Robustness to the quality of tracking mask: To understand the effect of the quality of tracking supervision, we experimented with a lower quality tracking mask by replacing segmentation masks with a bounding box, which is less accurate in terms of the shape of the object (Table 7d). The model obtained significant gain on PASCAL VOC over baseline (**+3.6**) and (-1.1) compared to our best model. Our model trained with unsupervised tracking mask still achieves comparable performance with the model using the supervised segmentation, which demonstrates its robustness to noises generated from unsupervised tracking.

Visual grounding and localization: We also visualize the grounding and localization ability of PreViTS-trained models finetuned on UCF-101 using Grad-CAM. Our model has a better grounding ability as compared to the baseline and focuses on foreground objects instead of background scenes (Figure 5). In Figure 6, we provide a query with two different segmentation corresponding to the different foreground objects. We feed the query and the key foreground into the PreViTS-trained model to compute the Grad-CAM attention heatmaps. Given the different key foreground, our model can localize the man and ball, respectively. At the same time, the attention heat map in the baseline is more spread out and cannot generate discriminative attention of the two objects. Even though PreViTS hasn't seen multi-object masks during pretraining, it is still able to localize multiple concepts discriminatively.

5. Conclusion

Limitations and potential impact: Our method has a few limitations. First, acquiring and utilizing unsupervised tracking requires additional computational resources. Also, since our current tracking method captures the most salient object in the video, we do not model multi-object interaction in the video, which is an interesting future work direction. Moreover, our pretraining datasets are relatively cleaner than random videos on YouTube. It is unknown if our method can generalize to the different genres such as news and gaming. Finally, our pretraining datasets may contain unintended societal, gender, racial, and other biases, whose effect was not examined in the current work.

Concluding remarks: We propose a visual self-supervised network that learns to localize foreground objects present in video data utilizing unsupervised tracking supervision. Experiments on various video downstream tasks show that guiding the model to focus on the foreground region is beneficial for accurate video representations self-supervised learning. Also, we demonstrate different properties of our learned features, which capture viewpoint, occlusion, illumination, and instance invariances. The result of our model shows better grounding ability with less background bias. We hope that our method leads to further research on robust, accurate and grounded visual representation learning from large-scale uncurated video data from the internet.

References

[1] Pulkit Agrawal, João Carreira, and Jitendra Malik. Learning to see by moving. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015.

[2] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020.

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016.

[5] Uta Buchler, Biagio Brattoli, and Bjorn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[6] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017.

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, 2020.

[8] 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 Conference on Artificial Intelligence*, 2021.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*, Proceedings of Machine Learning Research, 2020.

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *ArXiv preprint*, 2020.

[12] Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. Why can't I dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

[13] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *arXiv preprint arXiv:2101.07974*, 2021.

[14] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9726, 2022.

[15] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019.

[16] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2009.

[17] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *ArXiv preprint*, 2020.

[18] Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, and Weilin Huang. Cross-architecture self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19270–19279, 2022.

[19] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020.

[21] Olivier J. Hénaff. Data-efficient image recognition with contrastive predictive coding. In *Proc. of ICML*, Proceedings of Machine Learning Research, 2020.

[22] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. *ArXiv preprint*, 2021.

[23] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021.

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[25] Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. In *Ad-*

vances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[26] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, 2013.

[27] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *ArXiv preprint*, 2018.

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2011.

[29] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.

[30] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[31] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020.

[32] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*. Springer, 2016.

[33] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *ArXiv preprint*, 2021.

[34] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv preprint*, 2018.

[36] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*. Springer, 2016.

[37] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021.

[38] Deepak Pathak, Ross B. Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017.

[39] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[40] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

[41] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[42] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[43] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yi Xu, Xiang Wang, Mingqian Tang, Changxin Gao, Rong Jin, and Nong Sang. Learning from untrimmed videos: Self-supervised video representation learning with hierarchical consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13821–13831, 2022.

[44] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022.

[45] Nawid Sayed, Biagio Brattoli, and Björn Ommer. Cross and learn: Cross-modal self-supervision. In *German Conference on Pattern Recognition*. Springer, 2018.

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015.

[47] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.

[48] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

[50] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

[51] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018.

[52] 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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[53] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[54] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*. Springer, 2020.

[55] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015.

[56] Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.

[57] Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

[58] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

[59] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *Proc. of ICLR*, 2021.

[60] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *ArXiv preprint*, 2021.

[61] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.

[62] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. *ArXiv preprint*, 2021.

[63] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.

[64] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019.