Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency

Haiping Wu
McGill University, Mila

Xiaolong Wang
UC San Diego

Abstract

Recent works have advanced the performance of self-supervised representation learning by a large margin. The core among these methods is intra-image invariance learning. Two different transformations of one image instance are considered as a positive sample pair, where various tasks are designed to learn invariant representations by comparing the pair. Analogically, for video data, representations of frames from the same video are trained to be closer than frames from other videos, i.e. intra-video invariance. However, cross-video relation has barely been explored for visual representation learning. Unlike intra-video invariance, ground-truth labels of cross-video relation is usually unavailable without human labors. In this paper, we propose a novel contrastive learning method which explores the cross-video relation by using cycle-consistency for general image representation learning. This allows to collect positive sample pairs across different video instances, which we hypothesize will lead to higher-level semantics. We validate our method by transferring our image representation to multiple downstream tasks including visual object tracking, image classification, and action recognition. We show significant improvement over state-of-the-art contrastive learning methods. Project page is available at https://happywu.github.io/cycle_contrast_video.

1. Introduction

There has been a surge of recent interest in contrastive learning of visual representation [65, 29, 3, 28, 56, 9, 26, 40]. We have witnessed that contrastive learning outperforms supervised pre-training with large-scale human annotations in various visual recognition tasks [26, 9]. The key of this self-supervised task is to construct different views and transformations of the same instance, and learn the deep representation to be invariant to the view changes. To construct different views for forming positive image pairs in contrastive learning, the most common way is to use different data augmentations on the same instance (e.g. random cropping, image rotation, colorization).

However, simply performing artificial augmentation on single instance has shown its limitation in multiple applications [57, 66]. For example, Tian et al. [57] have performed detailed analysis on how different augmentations can affect different downstream visual recognition tasks. Going beyond single image, researchers have also looked into videos as the source for obtaining positive pairs of training images [48, 21, 60]. That is, two nearby frames in the same video can be taken as a natural augmentation in time for the same object instance. By training using temporal augmentation, the representation can learn viewpoint and deformation invariance. However, these approaches are still limited to find positive pairs and learning their similarity within a single instance.

In this paper, we propose to perform contrastive learning with positive image pairs sampled across different videos instead of the same video. We hypothesize this can potentially capture higher-level semantics and categorical information beyond low-level intra-instance invariances modeled by previous approaches. Specifically, given two image
frames \( I_i \) and \( I_j \) from a video, instead of directly using them as a positive pair for training \([48, 21]\), we will first "explain" frame \( I_i \) by composing frames from other videos that are similar to \( I_i \), then compare the composed frames to \( I_j \) for contrastive learning.

Assuming we have a neural network feature extractor to learn, we extract the feature representations for the image frame \( I_i \) as \( q_i \), and representations for frames from other videos as \( U = \{u_1, u_2, ..., u_m\} \). Given these representations, we compute the similarities between \( q_i \) and \( U \), and normalize them to a probability distribution. We use this probability distribution to re-weight and compose the features \( U \) as a new feature representation for frame \( I_i \) (frames that are more similar to \( I_i \) will have larger weight). We call this new feature as a soft nearest neighbor to \( I_i \).

We then form a positive pair of training data with this new representation and the feature of \( I_j \) (a different frame from the same video as \( I_i \)). As shown in Figure 1, this procedure goes through a cycle of starting from one frame \( I_i \) in a video, searching forward by matching frames from other videos, and retrieving frame \( I_j \) backward in the first video. We call this process Cycle-Consistent Contrastive Learning. Intuitively, enforcing such a cycle-consistency can explicitly push video frames with similar structure closer, thus leads to a natural clustering of semantics.

We perform the proposed self-supervised representation learning on unlabeled video dataset Random Related Video Views (R2V2) \([21]\) and transfer the learned representation to various downstream tasks including visual object tracking, image classification and action recognition. We stress our goal is to use the temporal signal to learn a general image-level representation for multiple applications beyond videos-level recognition tasks. We show significant improvements over multiple state-of-the-art approaches. We also conduct extensive ablation studies of different components and design choices of our method.

Our contributions include: (i) A novel cross-video cycle-consistent contrastive learning objective that explores cross-video relations, going beyond previous intra-image and intra-video invariant learning; (ii) The proposed loss enforces image representations from the same category (of similar visual structures) closer without explicitly generating pseudo labels; (iii) The learned image representation achieves significant improvement in multiple downstream tasks including object tracking, image classification and action recognition.

2. Related Work

Contrastive learning. The self-supervised contrastive learning methods \([23, 16, 65, 29, 44, 3, 28, 56, 78, 9, 26, 40, 10, 8]\) try to learn image representations under different transformations agree by forming positive and negative pairs, and make the representations of positive pairs have high similarity and negative pairs have low similarity. The typical way of generating positive pairs is performing artificial data augmentations on a single image instance. For example, Chen et al. \([9]\) introduce a contrastive learning baseline with different types of augmentations, including random cropping, resizing, color distortion, gaussian blur, etc. He et al. \([26]\) proposes MoCo which introduces a momentum network to encode a queue of a large number of egative samples for efficient learning. In this work, we build our model based on the MoCo framework. However, instead of learning with positive pairs by augmenting the same image, we propose a new objective which finds positive pairs of sample across videos for contrastive learning of image representations.

Self-supervised image representation learning from videos. Going beyond learning from a single image \([15, 13, 47, 14, 72, 19]\), video naturally offers temporal information and multi viewpoints for objects, which have been extensively utilized as self-supervisory signals in representation learning \([22, 1, 33, 60, 46, 45, 41, 61, 39, 7, 63, 35, 51]\). For example, Wang and Gupta \([60]\) use tracking to provide supervision signals that makes the feature representations of tracked patches similar. Recent works \([21, 48, 32, 68]\) further extend similarity learning between video frames under the contrastive learning framework, where the positive pairs for training are frames sampled from the same video. It has been shown image representations with viewpoint invariance can be learned. Our work is motivated by these previous works, and going beyond viewpoint invariance, learning using positive pairs across videos can potentially lead to image representations with higher-level semantics. While contrastive learning has also been applied to video representation learning with 3D ConvNets for action recognition \([35, 24, 36, 5, 50, 58, 25, 34, 42]\), we emphasize our work is focusing on learning a general image representation for multiple tasks beyond action recognition including visual tracking and image classification.

Cycle-consistency learning. Our work is influenced by cycle-consistency learning in different computer vision applications including 3D scene understanding \([31, 73, 20, 71]\), image alignment and translation \([74, 76, 75, 77]\), and space-time alignment in videos \([4, 62, 38, 17, 59, 32, 49, 37]\). For example, Wang et al. \([62]\) propose to perform forward and backward tracking in time to achieve a cycle-consistency for learning temporal correspondence. Dwibedi et al. \([17]\) formulates a temporal cycle consistency loss which aligns frames from one video to another between a pair of videos, and achieves good performance in video frame alignment tasks. Building on these two works, Pusenwalkam et al. \([49]\) propose to track object patches inside a video and align them across videos at the same time. While these results are encouraging, both approaches learning from video pairs \([17, 49]\) require human annota-
tors to provide ground-truth pairs (video-level) in training with a small scale of videos. In this paper, we propose to go beyond these restrictions, and apply cross-video cycle-consistency learning without any human annotations. These not only allows learning with a large-scale videos, but also generalizes our representations to multiple downstream vision tasks.

3. Cycle-Consistent Contrastive Learning

In this section, we first introduce contrastive learning with different forms of invariant learning targets. Then we propose our method with cross-video cycle-consistency learning.

3.1. Intra-Image and Intra-Video Invariance

The core of self-supervised contrastive learning [26, 9] is to learn representations which maximize the agreement between different views, augmentations of one image instance, and minimize the similarity between two different and unrelated instances at the same time. Most methods share the similar learning objective, which is to make the representations intra-image invariant. We describe the formulation of the objective as follows.

Given an query image $I_i$, the feature extractor encodes it to feature representations $q_i$ and $k_i$ under two different data augmentations. The intra-image invariance learning considers $q_i$ and $k_i$ as an a positive training pair and minimizes their representation distance, while maximizing the representation distance of $q_i$ and a set of negatives $U = \{u_1, u_2, ..., u_m\}$, which is a set of feature representations extracted from different images. The intra-image invariance contrastive learning loss function is defined as,

$$L_{\text{intra-image}} = -\log \frac{\exp (\text{sim}(q_i, k_i) / \tau)}{\sum_{u \in \{U, k_i\}} \exp (\text{sim}(q_i, u) / \tau)}, \quad (1)$$

where $\tau$ is the temperature constant and $\text{sim}(x, y) = x^\top y / \|x\|\|y\|$ is the cosine similarity between two feature vectors. The loss function tries to classify $q_i$ is similar to $k_i$ from the same image against features in $U$ from different images, achieving the intra-image invariance learning.

While intra-image invariance learning gives us good representations, augmentations of static images fail to capture viewpoint and deformation variations of instances (e.g., different viewpoints or gestures of one person) [48]. It is natural to resort to sequential video data that has variations of the same instance across time, extending the intra-image invariance learning to intra-video invariance learning. The intra-video invariance learning considers frames within the same video in a local time window as an invariant set to form positive pairs in training [52, 21, 48]. Similarly to Eq. 1, the intra-video invariance contrastive learning loss is defined as,

$$L_{\text{intra-video}} = -\log \frac{\exp (\text{sim}(q_i, k_j) / \tau)}{\sum_{u \in \{U, k_i\}} \exp (\text{sim}(q_i, u) / \tau)}, \quad (2)$$

where $k_j$ is the feature representation of image $I_j$, which is sampled from the same video of $I_i$. $L_{\text{intra-video}}$ tries to make the feature representations of images from the same video closer than images from different videos. When the sampled image $I_j$ is the same as $I_i$, $L_{\text{intra-video}}$ is identical to $L_{\text{intra-image}}$. We can see the intra-image invariance learning as a special case of the intra-video invariance learning.

3.2. Cross-Video Cycle-Consistency Objective

Both intra-image and intra-video invariance learning make the feature representations of the same instance (from the same image or video) closer. However, there is no explicit regularization on the distances between representations from the same class (or images of similar visual structures). For instance, the representations of different cat instances should be close and representations of frames from different videos on playing tennis should also be similar. In this section, we propose to find cross-video correspondence with cycle-consistency without using any ground-truth labels, and incorporate the correspondence in contrastive learning. The pipeline of the new proposed objective is shown in Figure 2 and we will introduce the formulations as following.

Our new objective consists of a forward and a backward nearest neighbor processes. Given a encoded query feature representation $q_i$ of $I_i$ from a video $V$, we first compute its nearest neighbor $\hat{q}_i$ in a candidate set $U$ containing frames from any videos. Then, we find the nearest neighbor of $\hat{q}_i$.
from the same video invariant set of \( I_k \) where \( \tau \) is the temperature and \( \text{sim}(x, y) = x^\top y / \| x \| \| y \| \) is the cosine similarity.

Given the soft nearest neighbor \( \hat{q}_i \), we assume if a representation is good for understanding the high-level semantics, \( \hat{q}_i \) should be in the invariance set of \( q_i \). Recall that we denote the invariance target of \( q_i \) as \( k_j \) (feature of a different frame in the same video) in intra-video invariance learning. We perform non-parametric classification as the backward step, and the cross-video cycle contrastive loss is defined as,

\[
\mathcal{L}_{\text{cycle}} = - \log \frac{\exp \left( \text{sim} \left( \hat{q}_i, k_j \right) / \tau \right)}{\sum_{u \in \{ U, k_j \}} \exp \left( \text{sim} \left( \hat{q}_i, u \right) / \tau \right)},
\]

where \( \lambda \) is a balancing factor of the two learning targets. We will provide ablation on \( \lambda \) in our experiments.

Algorithm 1 Pseudocode for \( \mathcal{L}_{\text{cycle}} \), PyTorch-like

```python
f_q: encoder + projection mlp
f_k: momentum encoder + projection mlp
mon: momentum parameter
queue: key dictionary (c x K)
m: size of neighbor set
t: temperature

\( f_k.params = f_q.params \)
for \( x_i, x_j \) in loader: # load a minibatch with n samples, each sample contains two randomly selected frames from one video
\( x_i, x_j = aug(x_i), aug(x_j) \) # random augmentation
q_i, k_j = f_q(x_i), f_k(x_j) # feature extraction, n x c
k_j = k_j.detach()

u = random.choice(queue, m) # randomly select neighbor set
r_queue = queue \ u # get remain elements of queue
a = softmax(cos_similarity(q_i, u)/t) # calculate similarity: n x m
soft_nn = matmul(a, u) # get soft nearest neighbor
logits = matmul(soft_nn, cat([k_j.T, r_queue], dim = 1))
labels = zeros(n)
cycle_loss = CrossEntropyLoss(logits/t, labels) # calculate cycle back classification loss
enqueue(queue, k_j)
dequeue(queue)
```

matmul: matrix multiplication; cat: concatenation.

Implementation details for neighbor set \( U \). In our experiments, we use two separate nearest neighbor sets for forward and backward steps for our objective \( \mathcal{L}_{\text{cycle}} \). The nearest neighbor set \( U \) in Eq. 3 is selected by randomly sampling from the current memory bank [26] at each training iteration. The remaining elements in the memory bank are used as negative candidates for finding the nearest neighbor backwards in Eq. 5. We provide the pseudo code of the \( \mathcal{L}_{\text{cycle}} \) in Alg. 1.

4. Experiments

In this section, we conduct experiments to perform unsupervised representation learning using the proposed learning objectives. We show that the learned representation transfer well to various downstream tasks. Then we design extensive ablation experiments to study the effectiveness of the proposed cross-video cycle-consistent learning method.

4.1. Experiments settings

Dataset. We perform unsupervised representation learning on Random Related Video Views (R2V2) [21] dataset, which is a large-scale diverse video frames collection. It has 2.7M videos, with 4 frames for each video. For smaller models and ablation studies, we use a subset of it for time efficient, which has 109k videos and 438k frames in total, which we refer as R2V2-S.

Network Architecture. ResNet-18 [27] is used as backbone for major ablation studies for its efficiency and ac-
4.2. Transfer to visual object tracking

We perform object tracking on the learned representations. SiamFC [6] is used as tracking method, which consists of one 1x1 convolution upon the pre-trained frozen representations. The training is performed on GOT-10k [30] dataset and we test it on OTB2015 [64]. The results are shown in Table 1. As we could see, when using ResNet-18 as backbone, our method outperforms previous unsupervised representation methods, as well as ImageNet supervised one, obtaining 65.6 precision, an 2.7 improvement over VINCE [21], which trains on the same dataset. Our method also surpasses previous contrastive learning methods when using ResNet-50 as backbone, and achieves 3.5 precision improvement over ImageNet supervised pretrain model. Note that we observed a performance drop when using intra-image objective alone when switching from ResNet-18 to ResNet-50. This also occurred in VINCE [21] which uses the same dataset for pretraining. We conjure that it is because there are only 4 frames per video, covering a long temporal window, making it hard to find correspondence within the same video, which is important for tracking task. Although our method is slightly inferior to SeCo [70]. However, SeCo utilizes a two-stage training strategy, where it first uses MoCo [26] method training on ImageNet dataset, and then trains Kinetics. Our method completely trains from scratch on R2V2, which makes SeCo is not directly comparable to ours.

In order to validate the effectiveness of our method, we conduct ablation study on different loss components with respect to the tracking performance, the results are shown in Table 2. It shows that our proposed \( L_{\text{cycle}} \) clearly improves methods considering only intra-image or intra-video constrative learning. Note that when training with only intra-image invariance without using temporal invariance in videos in R2V2 dataset, the performance for tracking will drop significantly, which is consistent with the results shown in VINCE [21].

| Method       | Backbone | Dataset     | OTB       | Success |
|--------------|----------|-------------|-----------|---------|
| Supervised   | ResNet-18| ImageNet    | 61.4      | 43.0    |
| SimSiam      | ResNet-18| ImageNet    | 58.8      | 42.9    |
| MoCo         | ResNet-18| ImageNet    | 62.0      | 47.0    |
| VINCE        | ResNet-18| R2V2        | 62.9      | 46.5    |
| Ours         | ResNet-18| R2V2-S      | 65.6      | 48.6    |

Table 1: Visual Object Tracking performance on OTB-100 compared with other unsupervised representation pretrain methods. SiamFC with one additional 1x1 convolution is added. Note that concurrent work SeCo* uses two-stage training strategy, where it uses MoCo pretrained on ImageNet as first stage.

### 4.3. Transfer to image classification

| Invariance | OTB       | Backbone | Success |
|------------|-----------|----------|---------|
| intra      |          | ResNet-18| 53.7    | 41.2    |
| video      |          | ResNet-18| 60.1    | 45.5    |
| cross      |          | ResNet-18| 65.6    | 48.6    |

Table 2: Ablation of different losses components (invariance target) for OTB-100 tracking on frozen features. Representations are pretrained on R2V2-S with ResNet-18, and R2V2 with ResNet-50. We clearly see that our proposed method with cross-video cycle consistency learning target achieves the best performance.
our transferring dataset for ablation study, which has 126k train images and 20k test images from 100 classes.

**Effect of different loss components.** In this part, we study the effect of using different losses to perform unsupervised representation learning and validate the performance on transferring to ImageNet classification. We study three different target of invariance learning, correspond to three loss functions, (a) \( L_{\text{intra-video}} \), (b) \( L_{\text{intra-video}} \), (c) \( L = L_{\text{intra-video}} + \lambda L_{\text{cycle}} \) (\( \lambda = 0.1 \)). The unsupervised representation training is performed on R2V2-S (ResNet-18), R2V2 (ResNet-50) dataset. The results are shown in Table 4.

As we could see, using a intra-video invariance learning target \( L_{\text{intra-video}} \) improves the ImageNet Top-1 accuracy from 33.0% to 33.1%, indicating that frames within the same video are natural views of the query image. Furthermore, our method, which adds \( L_{\text{cycle}} \) to perform cross-video cycle contrastive learning, further boosts the Top-1 accuracy to 34.4%, which is an absolute 1.4% improvement compared to intra-image invariance learning. With a deeper backbone ResNet-50, a similar trend is observed as the proposed full loss with \( L_{\text{cycle}} \) improves 1.8% Top-1 accuracy over intra-image objective. This shows that adding \( L_{\text{cycle}} \) makes the model explore cross-video relation, making features of visually similar instances across frames and videos closer.

| Invariance | Backbone | ImageNet Top-1 (%) |
|------------|----------|-------------------|
| intra image | intra video | cross video | ResNet-18 |
| ✓           |           | ✓               | 33.0       |
| ✓           |           | ✓               | 33.1       |
| ✓           |           | ✓               | 34.4       |
| ✓           |           | ✓               | 53.8       |
| ✓           |           | ✓               | 55.1       |
| ✓           |           | ✓               | 55.6       |

Table 4: Ablation of different losses components ImageNet classification on frozen features. Representations are pretrained on R2V2-S with ResNet-18 and R2V2 with ResNet-50. We clearly see that our proposed method with cross-video cycle consistency learning target achieves the best performance.

**Intra-image or Intra-video invariance are essential for good representation.** While exploring cross-image or cross-video information are helping learning better representation, it should be built on representation that is invariant of instance or videos. We conduct experiments that use \( L_{\text{cycle}} \) only to learn the representation. A random view (data augmentation) \( k_{++} \) of the query \( q_i \) in the neighbor set \( U \). The results are shown in Table 5. As we could see, when the neighbor size is small (e.g. 256), using \( L_{\text{cycle}} \) alone could learn representation that has better performance when transferring to linear classification on ImageNet-100, compared to using \( L_{\text{intra-video}} \). However, when the neighbor set is large, the performance drops largely. On the other hand, if \( k_{++} \) is not included in the neighbor set, \( U \), using a neighbor size of 256, and \( L_{\text{cycle}} \) alone would lead to 45.78% Top-1 accuracy on ImageNet-100, which is much worse. We safely draw that when neighbor size is small and a random view \( k_{++} \) of the query is included in the neighbor set, \( L_{\text{cycle}} \) degrades to \( L_{\text{intra-video}} \) by learning to make the similarity of \( q_i \) and \( k_{++} \) maximized. Thus, for making the model learn truly cross-video relation, excluding \( k_{++} \) from the neighbor set \( U \) is necessary. However, directly learning correspondence cross videos would be hard and shows worse results of 45.78% Top-1 accuracy on ImageNet-100. It is essential to add \( L_{\text{intra-video}} \) with \( L_{\text{cycle}} \) (i.e. using the full loss \( L \) in Eq.), boosting to 58.50 % Top-1 accuracy on ImageNet-100.

### How to choose neighbors?
We conduct experiments to study how the size of neighbor set \( U \) is affecting the performance. Full loss \( L \) is used to perform unsupervised representation learning on R2V2-S dataset and the size of \( U \) varies from 256 to 16384. Views of the current query are not added to the neighbor set. Then we evaluate the learned representation by conducting linear classification on frozen representation on ImageNet-100. The results are shown in Table 6. As we could see, increasing the neighbor size give us better representation on behalf of better ImageNet-100 Top-1 accuracy. This is expected as a large neighbor size of \( U \) would gives higher probability that the query could find correspondence across videos. The best Top-1 accuracy of 58.48% is achieved when using 16384 as neighbor size and we use this as default setting hereafter.

We also study that if using top-K nearest neighbors to perform the reconstruction in Eq. 3 would help representation learning. Top-K neighbors are selected by ranking the similarity \( u_{q_i,u} \) in the initial neighbor set, and the top K neighbors are chosen to construct the new neighbor set.

| Size of neighbor set \( U \) | 128 | 256 | 512 | 16384 |
|----------------------------|-----|-----|-----|-------|
| Acc Top-1 (%)              | 56.56 | 57.00 | 56.40 | 52.96 |
| Acc Top-5 (%)              | 83.10 | 83.06 | 82.64 | 79.46 |

Table 5: Ablation study. Accuracy on ImageNet-100 under linear classification protocol varying the size of the neighbor set \( U \). Representations are pretrained using \( L_{\text{cycle}} \) only on R2V2-S with ResNet-18. One view of the current query image is included in the neighbor set.
We study the effect of different losses on the task of video action recognition on UCF100 dataset. UCF100 has 13320 videos from 101 action categories. We train and test our model on split1 of UCF101. For simplicity, we directly use ResNet other than 3D convolution based methods. The video representation is obtained by averaging the frame representation from the video, and one fully-connected layer is used for predicting the action class on this video representation. Following [42], multiple clips are sampled from the video and the predictions of the clips are averaged for the final results. The results are shown in Table 9. As we could see, our method is able to achieve 76.8% top-1 accuracy on UCF101 with ResNet-18 as backbone, surpassing the previous best one of 68.2% with similar number of parameters.

We also list some results of methods with large 3D ConvNets models in Table 9. Our method is not directly comparable to these methods as we use 2D ConvNets, which not only have less parameters but also much less FLOPs compared to 3D ConvNets. We emphasize that our representation is able to solve multiple downstream tasks while the previous 3D ConvNets are designed only for action recognition. Note that, while DPC with large 3D-ResNet34 (32.6M) as backbone achieves 75.7%, our model with fewer parameters (11.7M), is able to surpass DPC with 3D-ResNet18 (14.2M) as backbone. Notably, when using ResNet-50 as backbone, our method is able to achieve 82.1%, surpassing other methods with larger 3D ConvNets models (e.g. MemDPC) as well. The results validate the effectiveness and transferability of the learned representations of our method.

### 4.4.2 Ablation Study

We study the effect of different losses on the task of video action recognition on UCF100 dataset. The pre-trained representations are fixed and a single fully-connected layer is added to predict the action class upon the averaged frame representations. The results are shown in Table 10. As we could see, for video action recognition task, it is beneficial to learn representations that are invariant within the same video, compared to invariant to the same image, rises the Top-1 accuracy from 45.1% to 48.4%, with ResNet-18 as backbone. However, it is also beneficial to make representations of similar videos closer other than separating them, as adding our cross-video cycle-contrastive learning loss $L_{cycle}$ further improves the Top-1 accuracy to 50.5%. Similarly, when using ResNet-50 as backbone, adding our $L_{cycle}$ achieves the best results of 67.1%. The results validate that our proposed cross-video cycle-contrastive learning target

| Size of neighbor set $U$ | 256 | 1024 | 4096 | 16384 |
|--------------------------|-----|------|------|-------|
| Acc Top-1 (%)            | 56.68 | 57.36 | 56.82 | **56.48** |
| Acc Top-5 (%)            | **84.04** | 83.78 | 83.52 | 83.50 |

Table 6: Ablation study. Accuracy on ImageNet-100 under linear classification protocol varying the size of the neighbor set $U$. Representations are pretrained using full loss $L$ on R2V2-S with ResNet-18.

| K          | 8    | 16   | 32   | 128  | 256  |
|------------|------|------|------|------|------|
| Acc Top-1 (%) | 57.52 | 57.40 | **57.64** | 57.62 | 56.98 |
| Acc Top-5 (%) | 83.02 | 82.84 | 83.62 | 83.76 | **83.82** |

Table 7: Ablation study. Accuracy on ImageNet-100 under linear classification protocol. Ablation on K when using top-K nearest neighbors to perform cross video cycle consistency learning. Representations are pretrained on R2V2-S with ResNet-18.

| $\lambda$ | 0.05 | 0.1  | 0.3  | 0.5  | 0.7  | 1.0  |
|-----------|------|------|------|------|------|------|
| Acc Top-1 (%) | 57.16 | **58.48** | 57.04 | 57.90 | 57.48 | 57.84 |
| Acc Top-5 (%) | **83.08** | 83.50 | 83.80 | 83.50 | 83.18 | **83.70** |

Table 8: Ablation results on loss balancing factor $\lambda$ in Eq. 6, Top-1 and Top-5 accuracy results of ImageNet-100 linear classification on frozen representations are shown. Representations are pretrained on R2V2-S with ResNet-18.
**Table 9:** Video action recognition accuracy comparison with other unsupervised representation methods on UCF101. We compare with methods using RGB modality. We mainly compare with models that has similar parameters as ours and list some large 3D ConvNet models (in gray) for reference.

| Method                  | Backbone                  | #Param | Dataset         | UCF101 |
|-------------------------|---------------------------|--------|-----------------|--------|
| Random Initialization [24] | 3D-ResNet18 | 14.2M  | -               | 46.5   |
| ImageNet Supervised Pretrained [53] | VGG-M-2048 | 353M  | -               | 73.0   |
| Shuffle & Learn [41] (227 × 227) | AlexNet       | 61M    | UCF101/HMDB51   | 50.2   |
| OPN [39]                     | AlexNet       | 61M    | UCF101          | 56.3   |
| DPC [24] (128 × 128)         | 3D-ResNet18 | 14.2M  | UCF101          | 60.6   |
| 3D-RotNet [35] (112 × 112)   | 3D-ResNet18-full | 33.6M  | Kinetics-400    | 62.9   |
| 3D-ST-Puzzle [36] (224 × 224) | 3D-ResNet18-full | 33.6M  | Kinetics-400    | 63.9   |
| SpeedNet [5] (224 × 224)     | 1D            | 12.1M  | Kinetics-400    | 66.7   |
| DPC [24] (128 × 128)         | 3D-ResNet18 | 14.2M  | Kinetics-400    | 68.2   |
| DPC [24] (224 × 224)         | 3D-ResNet34 | -      | Kinetics-400    | 75.7   |
| Video-Pace [58]             | R(2+1)D      | 33.3M  | Kinetics-400    | 77.1   |
| CBT [55] (112x112)           | S3D           | -      | Kinetics-400    | 79.5   |
| MemDPC [25]                  | R-2D3D       | 32.4M  | Kinetics-400    | 78.1   |
| Temporal-ssl [34]           | R(2+1)D      | 33.3M  | Kinetics-400    | 81.6   |
| VTHCL [69]                   | 3D-ResNet50  | 31.7M  | Kinetics-400    | 82.1   |

**Table 10:** Ablation of different losses components UCF action recognition learned on frozen features. Representations are pre-trained on R2V2 with ResNet-18 and ResNet-50.

| Invariance       | Backbone    | UCF101 Top-1 (%) |
|------------------|-------------|------------------|
| intra image      | ResNet-18   | 45.1             |
| intra video      | ResNet-18   | 48.4             |
| cross video      | ResNet-18   | 50.5             |
| intra image      | ResNet-50   | 62.2             |
| intra video      | ResNet-50   | 66.2             |
| cross video      | ResNet-50   | 67.1             |

**4.4.3 Nearest Neighbor evaluation**

To further validate our representation could learn cross-video information, we perform nearest neighbor retrieval experiments on the learned representation on UCF101 dataset on both frame level and clip level.

For frame retrieval experiments, following [7], 10 frames are sampled for each video. The representations of the frames from test set are used to find the nearest neighbors on the train set. For clip retrieval experiments, following [67], 10 clips per video are sampled. The clips extracted from the test set are used to find nearest neighbors on the train set. Cosine distance of representations is used as ranking criterion. If the class of the query sample appears in the class set of the k nearest neighbors, it is considered as a correct retrieval.

The results are shown in Table 11. Our model is able to surpass previous methods largely on both frame retrieval and clip retrieval experiments. Notably, our method with ResNet-18 as backbone has a superior accuracy when considering small k (e.g. k=1, 5, 10). For instance, our method achieves a top-1 accuracy of 45.8% on frame retrieval experiments, where previous best result from Buchler et al. [7] is 25.7%, which is an absolute 20.1% improvement. For clip retrieval, our method has a top-1 accuracy of 39.7%, an absolute 25.6% improvement compared to previous best of 14.1%. This indicates that our method is able to make the representations of similar frames/videos (short clips) closer, as they have higher probability belonging to the same class. Furthermore, our model is trained solely on R2V2 dataset and has never seen samples from UCF101 dataset, while other methods is trained on UCF101. The results indicate the transferability of our method. Video representations are obtained by averaging clip representations. We could see that even though our model does not have ground truth class to guide the representation learning, it manages to make video representations from the same class or having similar visual structures close.

Overall, we design various ablation experiments to study different components of the loss, including image recognition, video recognition, video retrieval, tracking. The results validate the effectiveness of our proposed cross-video cycle-contrastive learning target loss $L_{cycle}$.
| Methods                  | Training Data    | Top-1 | Top-5 | Top-10 | Top-20 | Top-50 |
|-------------------------|------------------|-------|-------|--------|--------|--------|
| **Frame retrieval:**    |                  |       |       |        |        |        |
| Jigsaw [43](CaffeNet)   | ImageNet         | 19.7  | 28.5  | 33.5   | 40.0   | 49.4   |
| OPN [39](CaffeNet)      | UCF101           | 19.9  | 28.7  | 34.0   | 40.6   | 51.6   |
| Buchler [7](CaffeNet)   | ImageNet+UCF101  | 25.7  | 36.2  | 42.2   | 49.2   | 59.5   |
| Ours (ResNet-18)        | R2V2             | 45.8  | 56.2  | 61.4   | 67.0   | 75.2   |
| Ours (ResNet-50)        | R2V2             | **52.6** | **63.3** | **68.1** | **73.3** | **80.6** |
| **Clip retrieval:**     |                  |       |       |        |        |        |
| Order [67] (C3D)        | UCF101           | 12.5  | 29.0  | 39.0   | 50.6   | 66.9   |
| Order [67] (R(2+1)D)   | UCF101           | 10.7  | 25.9  | 35.4   | 47.3   | 63.9   |
| Order [67] (R3D)        | UCF101           | 14.1  | 30.3  | 40.0   | 51.1   | 66.5   |
| SpeedNet [3] (S3D-G)    | Kinetics-400     | 13.0  | 28.1  | 37.5   | 49.5   | 65.0   |
| Ours (ResNet-18)        | R2V2             | 39.7  | 50.3  | 55.9   | 62.0   | 70.7   |
| Ours (ResNet-50)        | R2V2             | **46.8** | **56.7** | **62.1** | **67.6** | **75.1** |

Table 11: Frame retrieval and Clip retrieval results on **UCF101** compared to other unsupervised representation learning methods. Results of Jigsaw and OPN are taken from [7]. Our model largely surpasses previous methods and manages to do so without using UCF101 samples.

5. Conclusion

In this paper, we propose a cross-video cycle-consistent contrastive learning objective to perform self-supervised learning for image representations. The proposed method could learn to explore cross-video relations to make not only the representations of images within the same video closer, but also representations from different videos with similar visual structures close, without using ground truth class labels or generating pseudo labels. We perform self-supervised representation learning on unlabeled R2V2 video dataset, and show that the learned image representation transfer well to multiple downstream tasks including visual tracking, image classification and action recognition. Extensive ablation studies are conducted and validate the effectiveness of our proposed method. We hope our approach can open up an opportunity to utilize cross-instance paired data for learning general image representations.

References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In **ICCV**, 2015. 2
[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 8
[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019. 1, 2
[4] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In **ECCV**, 2018. 2
[5] 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. 2, 8, 9
[6] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 5
[7] 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)*, pages 770–786, 2018. 2, 8, 9
[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924, 2020. 2
[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 3
[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2
[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 5
[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5
[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2
[14] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 2
[15] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015. 2

[16] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014. 2

[17] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 2

[18] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. *arXiv preprint arXiv:2104.14558*, 2021. 8

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

[20] Clément Godard, Oisin Mc Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. 2017. 2

[21] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020. 1, 2, 3, 4, 5, 6

[22] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. *ICCV*, 2015. 2

[23] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[24] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 8

[25] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 2, 8

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 3, 4, 5, 6

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[28] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 1, 2

[29] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchilhon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 1, 2

[30] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[31] Qi-Xing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. In *Proceedings of the Eleventh Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, 2013. 2

[32] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *Advances in Neural Information Processing Systems*, pages 19545–19560, 2020. 2

[33] Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1413–1421, 2015. 2

[34] Simon Jenni, Givi Meishvili, and Paolo Favarlo. Video representation learning by recognizing temporal transformations. *arXiv preprint arXiv:2007.10730*, 2020. 2, 8

[35] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11587*, (2)(7):8, 2018. 2, 8

[36] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised spatiotemporal correspondence flow. *arXiv preprint arXiv:1905.00875*, 2020. 2

[37] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. *arXiv preprint arXiv:1905.00875*, 2019. 2

[38] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 2, 8, 9

[39] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. Cycle-contrast for self-supervised video representation learning. *arXiv preprint arXiv:2010.14810*, 2020. 2

[40] Ishan Misra and Lauren van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7607–7617, 2020. 1, 2

[41] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2, 8

[42] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. 2020. 2, 7, 8

[43] Mehdi Noroozi and Paolo Favarlo. Unsupervised learning of visual representations by solving jigsaw puzzles. In
[74] Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In CVPR, 2015.

[75] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 117–126, 2016.

[76] Xiaowei Zhou, Menglong Zhu, and Kostas Daniilidis. Multi-image matching via fast alternating minimization. In ICCV, 2015.

[77] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

[78] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In Proceedings of the IEEE International Conference on Computer Vision, pages 6002–6012, 2019.