Spatiotemporal Contrastive Video Representation Learning

Rui Qian[*][1,2,3] Tianjian Meng[*][1] Boqing Gong[1] Ming-Hsuan Yang[1]

Huisheng Wang[1] Serge Belongie[1,2,3] Yin Cui[1]

[1]Google Research [2]Cornell University [3]Cornell Tech

Abstract

We present a self-supervised Contrastive Video Representation Learning (CVRL) method to learn spatiotemporal visual representations from unlabeled videos. Our representations are learned using a contrastive loss, where two clips from the same short video are pulled together in the embedding space, while clips from different videos are pushed away. We find both spatial and temporal augmentation are crucial for video self-supervised learning. In particular, we propose a simple yet effective temporally consistent spatial augmentation method to impose strong spatial augmentations on each frame of a video clip while maintaining the temporal consistency across frames. CVRL shows superior performance on various tasks and datasets, *e.g.* for Kinetics-600 action recognition, a linear classifier trained on representations learned by CVRL achieves surprisingly **70.4%** top-1 accuracy with a 3D-ResNet50 backbone, significantly outperforming ImageNet supervised pre-training and SimCLR unsupervised pre-training, and greatly closing the gap between unsupervised and supervised video representation learning.

1 Introduction

Representation learning is of crucial importance in computer vision tasks, and a number of highly promising recent developments in this area have carried over successfully from the static image domain to the video domain. The temporal dimension of videos gives rise to key differences between them. However, self-supervised learning gravitates to different dimensions in images and videos, respectively. It is natural to engineer self-supervised learning signals along the temporal dimension in videos. Examples abound, including models for predicting the future [1–3], changing temporal sampling rates [4], and sorting video frames or clips [5–7]. Meanwhile, in the domain of static images, some recent work [8–11] that exploits the spatial dimensions has reported unprecedented performance on self-supervised image representation learning.

In this work, we show that apart from the commonly used temporal cues, self-supervised signals in the spatial subspace of videos also matters to video representation learning. Motivated by utilizing both spatial and temporal information, we build a self-supervised framework of contrastive video representation learning (CVRL) upon SimCLR [9] in view of its simplicity and compelling performance in the image domain. As illustrated in Figure 1, this framework learns both spatial and temporal information from videos by contrasting the similarity between a positive pair from the same video to those of negative pairs from different videos using the InfoNCE contrastive loss [12]. We further evaluate the learned video representations by **1)** linear evaluation protocol, **2)** the transfer learning setting via fine-tuning the entire network on other datasets following [6, 13–15] and **3)**

[*] The first two authors contributed equally. This work was conducted while Rui Qian worked at Google.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

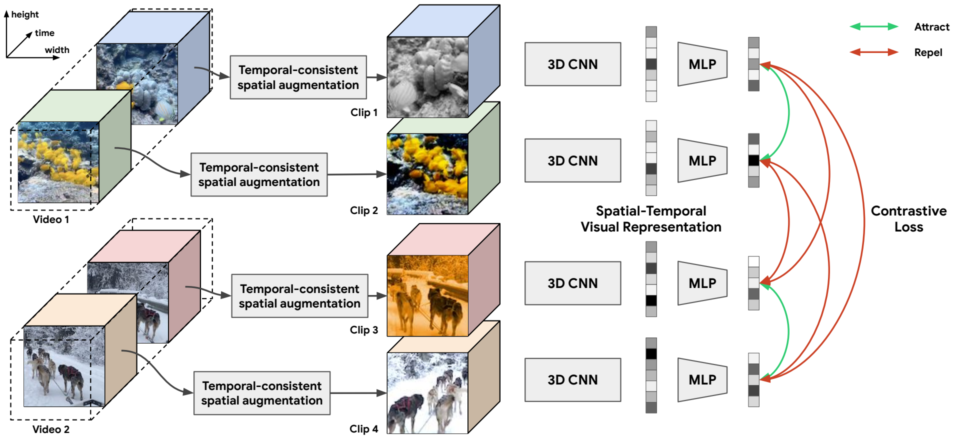
height
time
width
Video 1
Video 2
Temporal-consistent spatial augmentation
Temporal-consistent spatial augmentation
Temporal-consistent spatial augmentation
Temporal-consistent spatial augmentation
Clip 1
Clip 2
Clip 3
Clip 4
3D CNN
3D CNN
3D CNN
3D CNN
MLP
MLP
MLP
MLP
Spatial-Temporal Visual Representation
Attract
Repel
Contrastive Loss

Figure 1: **Overview of the proposed self-supervised contrastive video representation learning (CVRL) framework.** From a short video, we randomly sample 2 clips with the same length. We then apply a temporally consistent spatial augmentation to each of the video clips and feed it to a 3D backbone with an MLP head. The contrastive loss is used to train the network to attract clips from the same video and repel clips from different videos in the embedding space.

the semi-supervised learning setting via fine-tuning the entire network following [8–10]. We next summarize our main findings.

Our representation shows superior peformance. To the best of our knowledge, on the linear evaluation protocol, all state-of-the-art methods on self-supervised learning for videos report lower accuracy than 3D inflated networks pre-trained on ImageNet [4, 16]. For the metric of top-1 accuracy on Kinetics-400 [17] as shown in Table 1, CVRL achieves 75% relative improvement compared with the state-of-the-art method and 24% relative improvement compared with ImageNet pre-trained networks. For the transfer learning setting, CVRL achieves 18% relative improvement on UCF-101 [18] and 59% relative improvement on HMDB-51 [19] compared with the state-of-the-art method, as shown in Table 2. For semi-supervised setting in Table 3, CVRL surpasses all other baselines especially when there is only 1% labeled data for fine-tuning, indicating that the advantage of our self-learned CVRL is more profound with limited labeled data.

The self-supervised signals in the spatial subspace of videos do matter to video representations. Spatial augmentations outperform temporal augmentations by 15.6% top-1 accuracy in the linear evaluation protocol. Combining the spatial augmentations with temporal augmentations yields 8.8% gain over spatial augmentations only.

The temporal consistency in spatial augmentation is important. We propose a temporally consistent spatial augmentation to keep the motion cues across frames. More detail would be covered in Section 3. By adding the temporal consistency, a further improvement of 9.6% can be obtained.

2 Related Work

Self-supervised video representation learning. It is natural to exploit the temporal dimension in self-supervised video representation learning. Some early work predicts the future on top of frame-wise representations [1]. More recent work learns from predicting motion and appearance statistics [13] or encodings [2, 3]. Another common approach is sorting frames or video clips [5–7, 20] along the temporal dimension. Yang *et al*. learn by maintaining consistent representations of different sampling rates [4]. Furthermore, videos can often supply multi-modal signals for cross-modality supervision, such as geometric cues [21], speech or language [22], and audio [23].

Self-supervised image representation learning. Some early work learns visual representations from unlabeled images via manually specified pretext tasks, for instance, auto-encoding [24–26], relative patch location [27], jigsaw puzzles [28], and image rotations [29]. Recently, contrastive learning framework [8–10, 12, 30], which maintains relative consistency between the representations of an image and its augmented view, achieved great success.

3 Methodology

3.1 Video Representation Learning Framework

We build our self-supervised framework of contrastive video representation learning (CVRL) upon SimCLR [9] in view of its simplicity and compelling performance on learning image representations. Figure 1 shows an overview of our framework.

The core of this framework is an InfoNCE contrastive loss [12] applied on features extracted from augmented videos. Suppose we sample N videos per mini-batch and augment them, resulting in $2N$ videos. Denote by $\boldsymbol{z}_i, \boldsymbol{z}'_i$ the encoded representations of the two augmented versions of the i-th input video. The InfoNCE contrastive loss is defined as $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i$ and

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}'_i)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}, \tag{1}$$

where $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$ is the inner product between two vectors after normalizing them onto a unit sphere, $\mathbf{1}_{[.]}$ is an indicator excluding from the denominator the self-similarity of the encoded video $\boldsymbol{z}_i$, and $\tau > 0$ is a temperature parameter. The loss allows the positive pair $(\boldsymbol{z}_i, \boldsymbol{z}'_i)$ to attract mutually while they repel the other items in the mini-batch. We set $\tau = 0.1$ in all experiments.

We construct other components of the framework as follows: (1) an encoder neural network maps an input video clip to its representation $\boldsymbol{z}$, (2) spatiotemporal augmentations construct the positive pairs $(\boldsymbol{z}_i, \boldsymbol{z}'_i)$ and the properties they induce, and (3) methods to evaluate the learned representations.

3.2 Video Encoder

We encode a video sequence using 3D-ResNets [31] as backbones. We expand the original 2D convolution kernels to 3D to capture spatiotemporal information in videos. The design of our 3D-ResNets mainly follows the "slow" pathway of the SlowFast network [32] with two modifications: (1) the temporal stride of 2 in the data layer, and (2) the temporal kernel size of 5 and stride of 2 in the first convolution layer. We also take as input a higher temporal resolution of 16 frames.

3.3 Temporal Augmentation

We impose on the video representations an invariance-to-time property by a temporal augmentation within a short duration. To this end, we apply a simple temporal augmentation method by sampling two clips with different start frames randomly from an input video. We regard these two video clips as a positive pair, inducing the short-time invariance constraint on the learned representations.

3.4 Temporally Consistent Spatial Augmentation

Spatial augmentation is widely used in both supervised and unsupervised learning for images. In the video domain, a natural strategy is to utilize existing image-based spatial augmentation methods to the video frames one by one. However, this method could break the motion cues across frames since spatial augmentation methods often contain some randomness such as random cropping, color jittering and blurring as important ways to strengthen their effectiveness. Therefore, we propose a simple yet effective approach to address this issue, by making the spatial augmentations consistent along the temporal dimension. We adopt the spatial augmentation sequence of $[RandomResizedCrop, RandomFlip, ColorJitter, GaussianBlur]$ with temporal consistency.

4 Experiments

Comparison baselines. We compare our CVRL method with two baselines: (1) inflating ImageNet pre-trained ResNets to our 3D ResNets by duplicating the temporal dimension, called ImageNet inflated, and (2) SimCLR inflated by inflating ResNets pre-trained with SimCLR. In addition, the supervised learning serves as an upper bound of our method.

Self-supervised pre-training. We conduct self-supervised pre-training on Kinetics-400 [17] and Kinetics-600 [33] for 800 epochs with an initial learning rate of 0.32 and batch size of 1024. We use synchronized batch normalization to avoid information leakage [9].

Linear evaluation. We train a linear classifier on top of the fixed pre-trained weights in the backbone network. During training, we sample a 32-frame (2 temporal stride) clip from each video to train the linear classifier for 100 epochs with an initial learning rate of 32. As shown in Table 1, CVRL has superior performance on both datasets and greatly closes the gap to the supervised learning.

Transfer learning. Following the practice in [3, 4, 6, 7, 13–15, 34, 35], we use pre-trained weights on Kinetics-400 [17] to initialize the network and fine-tune all layers on UCF-101 [36] and HMDB-51 [19] datasets. As shown in Table 2, CVRL outpeforms state-of-the-art methods significantly.

Semi-supervised learning. We sample 1% and 10% videos in the training set to fine-tune the entire pre-trained network with an initial learning rate of 0.2. The evaluation set remains the same. As shown in Table 3, CVRL surpasses all other baselines, especially with only 1% labeled training data.

Ablation study on data augmentation. We verify the effectiveness of our data augmentation strategy with self-supervised pre-training and linear evaluation on Kinetics-400. As shown in Table 4, the proposed temporally consistent spatial augmentation is essential to obtain good performance.

Method	Dataset	Top-1 Acc.
VINCE [16]	Kinetics-400	36.2
VTHCL [4]	Kinetics-400	37.8
SimCLR	Kinetics-400	46.9
ImageNet	Kinetics-400	53.5
CVRL	Kinetics-400	**66.1**
Supervised	Kinetics-400	75.8
SimCLR	Kinetics-600	48.0
ImageNet	Kinetics-600	54.7
CVRL	Kinetics-600	**70.4**
Supervised	Kinetics-600	78.5

Table 1: **Linear evaluation results.**

Method	Top-1 Acc.	
	UCF-101	HMDB-51
MotionPred [13]	61.2	33.4
3D-RotNet [14]	62.9	33.7
ST-Puzzle [6]	65.8	33.7
SpeedNet [15]	66.7	43.7
ClipOrder [7]	72.4	30.9
DPC [3]	75.7	35.7
PacePred [34]	77.1	36.6
MemDPC [35]	78.1	41.2
CVRL	**92.1**	**65.4**

Table 2: **Transfer learning results.**

Method	Backbone	Kinetics-400 Top-1 Acc. (Δ *vs*. Sup.)		Kinetics-600 Top-1 Acc. (Δ *vs*. Sup.)	
		Label fraction		Label fraction	
		1%	10%	1%	10%
Supervised learning	3D-R50	3.2	39.6	4.3	45.3
ImageNet inflated	3D-R50	16.0 (12.8↑)	49.1 (9.5↑)	17.3 (13.0↑)	52.6 (7.3↑)
SimCLR inflated	3D-R50	18.6 (15.4↑)	46.5 (6.9↑)	19.7 (15.4↑)	48.3 (3.0↑)
CVRL	3D-R50	**35.1 (31.9↑)**	**58.1 (18.5↑)**	**36.7 (32.4↑)**	**56.1 (10.8↑)**

Table 3: **Semi-supervised learning results on Kinetics-400 and Kinetics-600.**

Augmentation	TA	SA	TA + SA	TA + SA + TC (proposed)
Top-1 Acc.	24.8	40.4	49.2	58.8

Table 4: **Ablation study on data augmentation.** All experiments are based on 100 epochs of self-supervised pre-training on Kinetics-400. TA indicates Temporal Augmentation. SA indicates Spatial Augmentation. TC indicates Temporal Consistency. (TA + SA + TC) is our proposed method.

5 Conclusion

This work presents a contrastive video representation learning (CVRL) framework for learning spatiotemporal representations from unlabeled videos. Extensive experiments on Kinetics-400, Kinetics-600, UCF-101 and HMDB-51 demonstrate promising results.

References

[1] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015.

[2] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[3] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV Workshops*, 2019.

[4] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*, 2020.

[5] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017.

[6] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019.

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

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

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

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

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

[13] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, 2019.

[14] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2(7):8, 2018.

[15] 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.

[16] 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.

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

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

[19] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[20] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, 2017.

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

[22] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019.

[23] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, 2018.

[24] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[25] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[26] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017.

[27] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

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

[30] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, 2019.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[32] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[33] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[34] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. *arXiv preprint arXiv:2008.05861*, 2020.

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

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