ViewCLR: Learning Self-supervised Video Representation for Unseen Viewpoints Supplementary

Srijan Das, Michael S. Ryoo

Method	Alignment Error
ViewCLR	**0.28**
ViewCLR-$\mathcal{L}_{3D}$	0.45
ViewCLR-$\mathcal{L}_{Adv}$	0.32
ViewCLR-$\mathcal{L}_{Adv}$-$\mathcal{L}_{3D}$	0.59

Table 1: Alignment Error of samples across different classes on NTU-60 (CVS3 protocol).

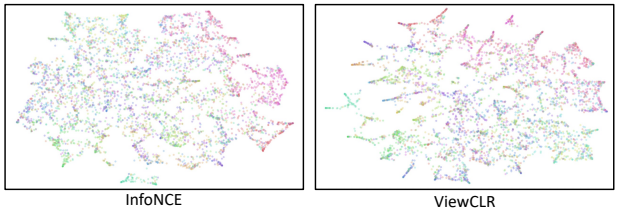

Figure 1: t-SNE representation of the test examples captured from unseen viewpoint for InfoNCE (at left) and ViewCLR (at right) models.

1. More Implementation Details

Training/Testing specification for downstream fine-tuning on NTU-60, NTU-120 and NUCLA. At training, we apply the same data augmentation as in the pre-training stage mentioned in section 4.2., except for Gaussian blurring. The model is trained with similar optimization configuration as in the pre-training stage for 500 epochs. At inference, we perform spatially fully convolutional inference on videos by applying ten crops (center crop and 4 corners with horizontal flipping) and temporally take clips with overlapping moving windows. The final prediction is the average *softmax* scores of all the clips.

ViewCLR adaptation with ρ-BYOL. BYOL can be viewed as a non-contrastive SSL method that does not use negative samples, but an extra predictor MLP is placed on top of encoder $f(\cdot)$. BYOL minimizes the negative cosine similarity between the feature computed by the predictor and the feature computed by the momentum updated version of encoder $f(\cdot)$. In our ViewCLR variant, we still retain a memory bank Queue1 in order to compute the 3D loss $\mathcal{L}_{3D}$ by computing the Top-1 nearest neighbor 3D world representation in Queue2 as explained in section 3.2 of the main paper. We follow the implementation specifics in [1] to train this variant of ViewCLR. As suggested in [1], we use a temporal persistency over $\rho = 4$.

2. Quantitative Evaluation of the Learned Camera Matrix

We quantify the effectiveness of the learned camera parameters by computing the alignment error of the 2D Projections within their action class in Table 1 on the test data of NTU-60 dataset. We chose to evaluate this metric on CVS3 protocol (of NTU-60) due to its challenging scenarios in terms of unseen viewpoints. The alignment error is computed by $\frac{1}{C^2}\sum_{k=1}^{C}\sum_{i,j=1}^{C}\text{dist}(P_k^i - P_k^j)$ where $i \neq j$, C is the # of classes and P_k is the center of the projected representation in world coordinate system. dist() is a distance measure between two coordinates, here we use Euclidean distance. ViewCLR with all its components achieve the lowest alignment error and hence shows its discriminative power in the feature space.

3. More Qualitative Visualization

In Figure 1, we provide a t-SNE [6] visualization of the samples captured from unseen camera viewpoints (NTU-60; CVS3 protocol) produced by the InfoNCE and ViewCLR models. These models are trained following the linear probe evaluation, hence the encoders are frozen with the pre-trained weights. It can be observed that ViewCLR better discriminates the action classes in the feature space compared to the traditional InfoNCE model substantiating the importance of the ViewCLR type pre-training.

In Fig. 2, we provide embedding based nearest neighbor retrievals. From the retrieval visualization, we find that ViewCLR often fails for similar actions like *put on a hat* and *take off hat*. This is because for such similar action representations, the 3D loss is guided by 3D latent representation of different action class. We also quantify the effectiveness of

Figure 2: Nearest Neighbor Retrieval with ViewCLR representations. The left side is the query video and the right side are the top-3 nearest neighbors from NTU-60 dataset.

the learned camera parameters by computing an alignment metric in the supplementary.

4. Regularization effect of ViewCLR

In Fig. 3, we provide (A) a plot of $(\mathcal{N}+1)$-way accuracy of the pretext task of the two models, ViewCLR and the MoCo model trained with InfoNCE loss, and (2) the training losses of the two aforementioned models. Note that ViewCLR is trained with InfoNCE loss for first 300 epochs as illustrated in Figure 3. We observe that the $(\mathcal{N}+1)$-way accuracy of the ViewCLR model while training on mixed contrastive loss is lower at times than that of the InfoNCE model (at the left of the figure). However, the performance gain of the ViewCLR model on downstream action classification task shows the regularizing capability of using view-generator. Meanwhile, we observe a disparity between the training losses (at right of the figure) in both the models ViewCLR and InfoNCE. This is owing to the hardness of the pretext task which can be directly correlated with the difficulty of the view-invariant transformation, via the data generator. This regularizing capability of ViewCLR is mainly obtained from the mixup strategy introduced in the MoCo model to infuse the view-invariant representation of the videos [3].

5. Model Architecture

Figure 4 provides a full illustration of ViewCLR presented in the main paper. $\mathcal{L}_R$ denotes the reconstruction loss incurred at the output of the view-generator to squeeze the dimension of the world latent feature F^W.

We also confirm the robustness of ViewCLR where it improves the baseline (MoCo counterpart) with different frame lengths (at left) and with different backbones (at right) in Table 2.

Method	Frame Length (t)		
	16	32	64
Baseline	72.5	73.8	73.9
ViewCLR	**74.9**	**75.8**	**75.8**

Method	**Backbone**		
	S3D	I3D	R3D
Baseline	73.8	73.4	71.1
ViewCLR	**75.8**	**75.6**	**72.9**

Table 2: Fine-tuning performance of ViewCLR w.r.t. Baseline MoCo for different frame length (at left) and for different backbones (at right) on NTU-60 dataset (CVS3 protocol).

6. Limitations

In this work, although we aim at learning video representation that generalizes in the wild to different camera viewpoints, our experiments are limited to videos captured in the indoors. This is owing to the scarcity of huge video datasets posing cross-view challenges. On one hand, large datasets like Kinetics [2] captured from the web mostly contain videos with viewpoint bias. On the other hand, multi-camera datasets like NTU-RGB+D [5], and MLB [4] are either constrained to indoors or contains less training data. Thus, we also encourage the vision community to pursue research towards learning domain agnostic representation of actions from web videos.

References

[1] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308, 2021.

[2] 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.

[3] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. In *ICLR*, 2021.

[4] AJ Piergiovanni and Michael Ryoo. Learning multimodal representations for unseen activities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[5] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

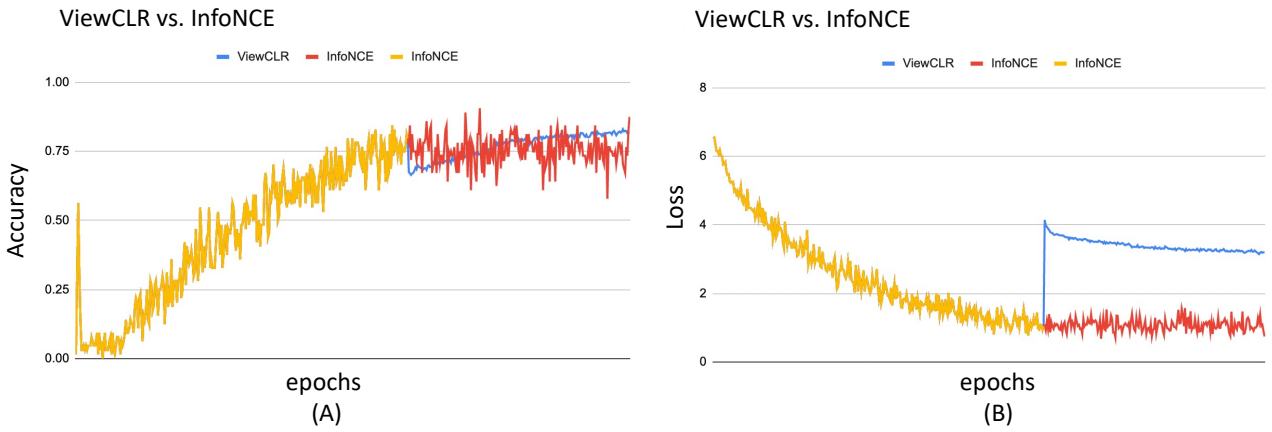

Figure 3: Training on NTU-60 dataset. At left, we provide the $(\mathcal{N}+1)$-way accuracy on the pretext task while learning contrastive representation. At right, we present the training loss of ViewCLR and InfoNCE models. ViewCLR models are initially pre-trained with InfoNCE loss for first 300 epochs.

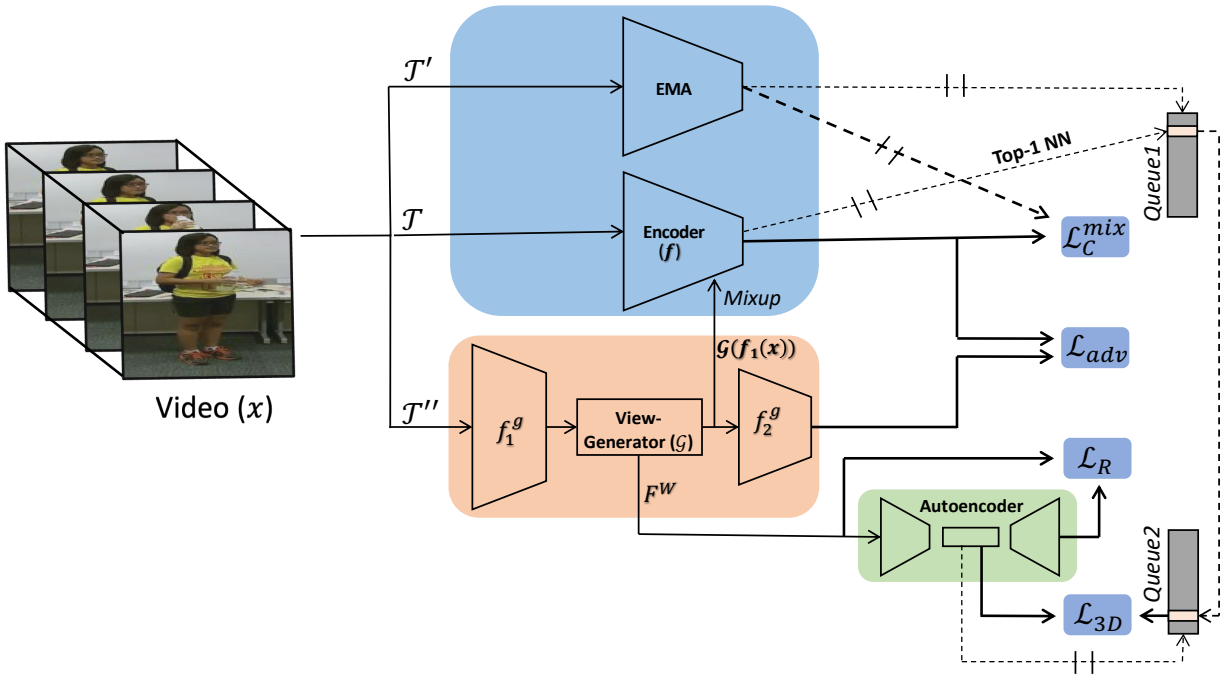

Figure 4: A full illustration of ViewCLR presented in Fig. 2 (right most) of the main paper.