Supplementary Material for
"PreViTS: Contrastive PREtraining with VIdeo Tracking Supervision"

This appendix is organized as follows:
1. Implementation details.
2. Additional experimental results.
3. Experiment details.
4. Additional qualitative results.

1. Implementation details

Image model is from MoCo, video model is from RSPNet. For experiments with the image model, we use the ResNet-50 backbone and sample one frame with 224 × 224 spatial sizes for each clip. For experiments with the video model, we use an S3D-g [15] backbone and sample 16 continuous frames with 224 × 224 spatial sizes for each clip. We perform standard data augmentation on clips, including random Gaussian blur, and random color jitter [2]. To compare with other baseline methods, we also trained on R(2+1)D[13], and C3D[12] backbone following [1]. We followed [1] to train our model with 200 epochs with SGD and a batch size of 256. We apply a cosine learning rate scheduler with an LR of 0.03 for the image model and 0.5 for the video model. Following He *et al.* [7], we set $\tau = 0.07$, $K = 65535$, $\gamma = 0.15$, $\mu = 0.3$, $\lambda = 3$. The training time is two days for pretraining VGG-Sound and three days for pretraining on Kinetics. For both image and video tasks, we compare with the following baselines: (1) **Random Init** of weights without pretraining, (2) **MoCo/RSPNet** to demonstrate standard self-supervised model performance for image (MoCo) and video (RSPNet), (3) **MoCo/RSPNet + Tracking Constrained Sampling** to evaluate our unsupervised tracking-based spatial-temporal sampling strategy.

2. Additional experimental results

Generalize to image recognition tasks. We evaluate our learned features on four downstream image recognition tasks: **(a)** PASCAL VOC [6] linear classification, **(b)** ImageNet-1k [4, 10] linear classification, **(c)** PASCAL VOC object detection, and **(d)** COCO [8] instance segmentation. Following [5, 11], for **(a, b)**, we perform linear classification by using the SSL model as a frozen feature extractor and training a classifier on top. For **(c, d)**, we use the SSL model as weight initialization for fine-tuning on the labeled datasets. Detailed experimental settings can be found in the supplementary. Our results in Table 1 show that training PreViTS outperforms baseline MoCo training on all tasks, obtaining robust gains in VOC and ImageNet classification, along with VOC detection and COCO instance classification. Notably, the performance gains when pretraining on VGG-Sound are larger as compared to those on Kinetics-400, even though Kinetics-400 is 20% larger in terms of the number of videos. We speculate that due to VGG-Sound containing a more diverse collection of objects as compared to Kinetics-400, which is primarily human action-centric, VGG-Sound benefits more from being able to learn object-focused representations when training with PreViTS. The performance improvement over baseline is especially large on the VOC detection task, aided by the improved ability to localize objects during pretraining. Finally, while it is typically challenging to obtain comparable performance to supervised ImageNet pretraining using video SSL pretraining on image recognition tasks [9], due to the larger domain shift, MoCo models trained with PreViTS still obtain comparable or better performance to ImageNet-fully supervised training on VOC detection and COCO instance segmentation tasks.

Video Backgrounds Challenge (mini-Kinetics). In addition to the video backgrounds challenge, we also evaluate robustness to background signal on the mini-Kinetics dataset [3], a subset of Kinetics-400 designed to study if video classification models depend on the background signal for scene classification. This dataset contains foreground bounding boxes computed by a person detection model. We utilize the bounding boxes to mask the foreground object to analyze if the model depended on scene features when performing action classification. The model with PreViTS achieved an accuracy of 55.24% in the Original setting compared to 47.18% for the baseline RSPNet. When the foreground was masked (No-FG), the accuracy for PreViTS drops by 6.9%, as compared to a drop of 2.71% for the baseline model, indicating that the PreViTS-trained model relies less on the background signal.

Computational resource compared to baseline. Obtaining tracking for a dataset is a fixed, one-time computational

Method	Dataset	VOC07 clf.	IN-1k clf.	PASCAL VOC Detection			COCO Instance Segmentation					
		mAP	Top-1 acc.	AP^{bbox}_{all}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{bbox}_{all}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}_{all}	AP^{mask}_{50}	AP^{mask}_{75}
1) Random Init		–	–	33.8	60.2	33.1	36.7	56.7	40.0	33.7	53.8	35.9
2) ImageNet Fully Sup		–	–	53.5	81.3	59.1	38.9	59.6	42.7	35.4	56.5	38.1
3) MoCo	K400	69.3	47.3	50.6	78.0	55.1	40.5	58.9	41.9	35.1	55.6	37.3
4) + Tracking Con. Sampling	K400	$70.4_{+1.1}$	$48.2_{+0.9}$	$51.2_{+0.6}$	$78.4_{+0.4}$	$56.1_{+1.0}$	$40.8_{+0.3}$	$59.5_{+0.6}$	$42.6_{+0.7}$	$35.8_{+0.7}$	$56.8_{+1.2}$	$38.3_{+1.0}$
5) + PreViTS	K400	$71.2_{+1.9}$	$48.6_{+1.3}$	$51.8_{+1.2}$	$78.3_{+0.3}$	$56.0_{+0.9}$	$41.0_{+0.5}$	$59.4_{+0.5}$	$42.8_{+0.9}$	$35.6_{+0.5}$	$57.2_{+1.6}$	$38.4_{+1.1}$
6) MoCo	VGG Sound	68.3	46.9	48.3	76.5	52.6	38.4	58.7	41.9	35.0	55.8	37.2
7) + Tracking Con. Sampling	VGG Sound	70.3_{+2}	$48.1_{+1.2}$	$49.0_{+0.7}$	$77.1_{+0.6}$	$52.7_{+0.1}$	$38.3_{-0.1}$	$58.7_{+0.0}$	$41.7_{-0.2}$	$35.0_{+0.0}$	$55.9_{+0.1}$	$37.6_{+0.4}$
8) + PreViTS	VGG Sound	$73.0_{+4.7}$	$50.6_{+3.7}$	$52.5_{+4.2}$	$78.7_{+2.2}$	$55.1_{+2.5}$	$39.4_{+1.0}$	$59.8_{+1.1}$	$43.0_{+1.1}$	$35.7_{+0.7}$	$56.8_{+1.0}$	$38.2_{+1.0}$

Table 1: **Transfer Learning on Image Downstream Tasks:** On tasks using linear probes (VOC and ImageNet classification) and finetuning (VOC Detection, COCO Segmentation), PreViTS outperforms baseline MoCo when evaluated on models pretrained on VGG-Sound and Kinetics-400. We color the difference ≥ 0.5 to show improvement over the baseline MoCo models (row 3 and 6).

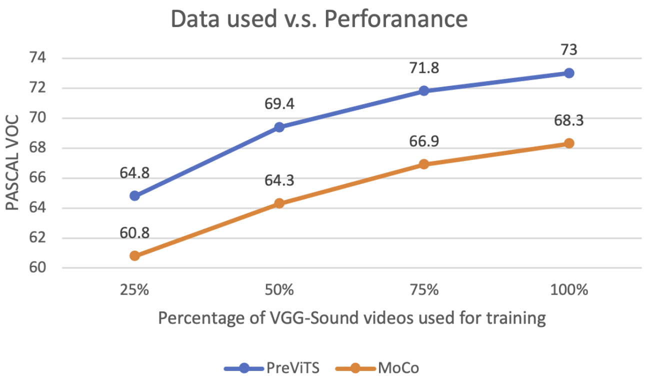


Figure 1: Percentage of VGG-Sound videos used for training.

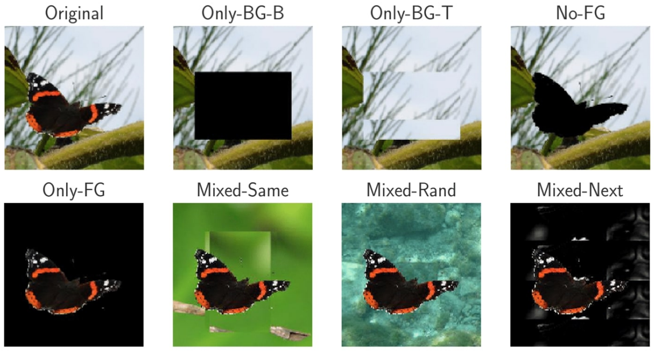


Figure 2: Image Background Challenge Settings

cost. During training, PreViTS only needs 1.3x GPU memory and training time due to the extra forward pass for the foreground key and query to compute Grad-CAM. PreViTS is also efficient, it outperforms baseline with only half of the training data (VGG-Sound), i.e., 65% of its training time in Figure 1.

Method Complexity of PreViTS. While PreViTS contains several components, it is not sensitive to their hyperparameters and design choices. To test sensitivity, we randomly chose a combination of parameters μ, λ, using the setting in Tab. 1(8) in the main paper and obtained **+4.32** VOC07 mAP over the baseline, only lower by **-0.38** than our best model.

Evidence for lack of proper supervisory signal in current SSL approaches. As visualized in Fig. 1(d) in the main paper, simply applying contrastive loss may lead to learning background correlation when the backgrounds are similar. Moreover, through a study using supervised segmentation on VGGSound, we found that traditional SSL approaches sample different concepts as positive pairs 27% of the time, while only 7% with our spatio-temporal sampling strategy. This indicates our strategy can acquire a cleaner supervisory signal.

3. Experiment details

Image Backgrounds Challenge. The settings of different scenarios of backgrounds are shown in Figure 2. The figure is from [14].

Code of the paper. We will release our code by the time when the paper is published.

4. Additional qualitative results

We include more visualizations for UCF-101 action recognition in Figure 3, Video Backgrounds Challenge in Figure 4, and DAVIS video object segmentation in Figure 5 and 6.

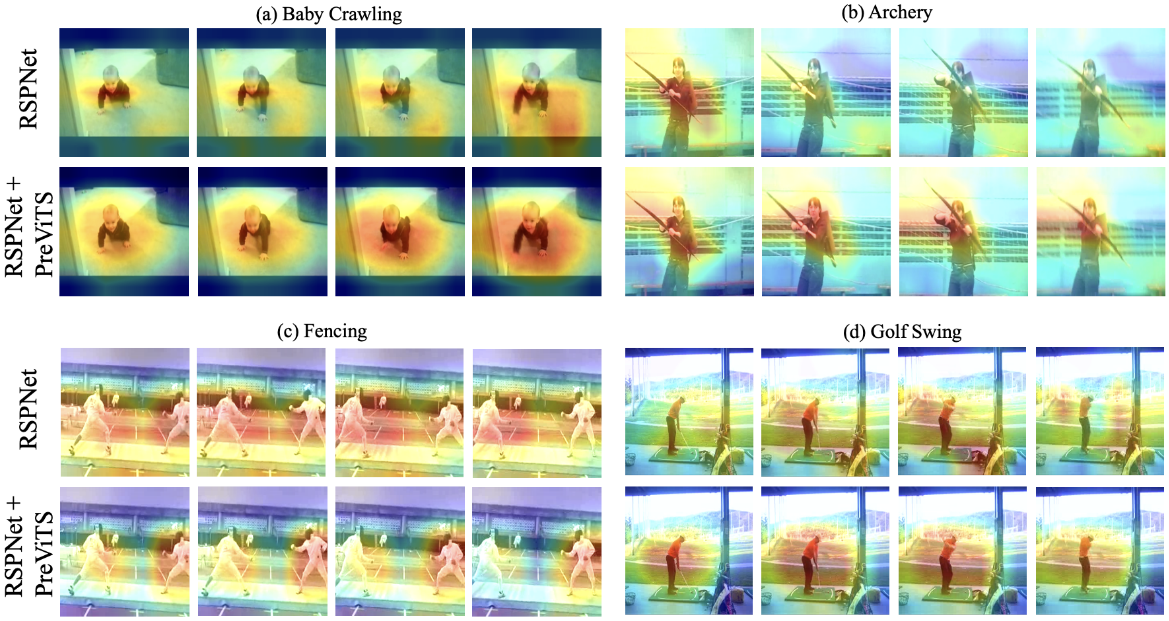


Figure 3: Grad-CAM Visualization for UCF-101 Action Classification.

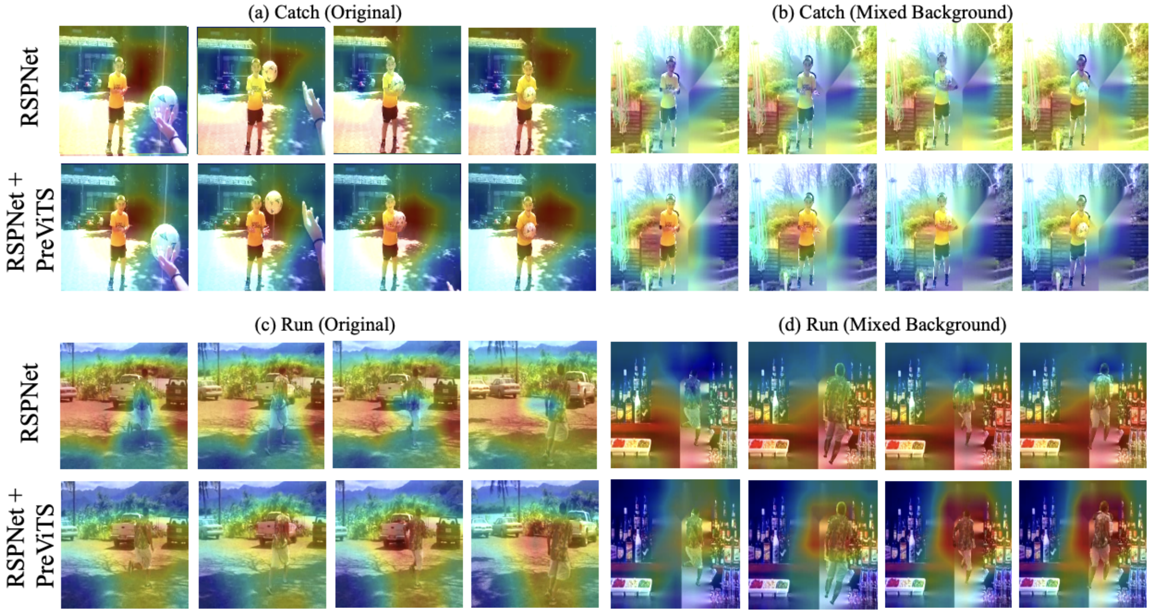
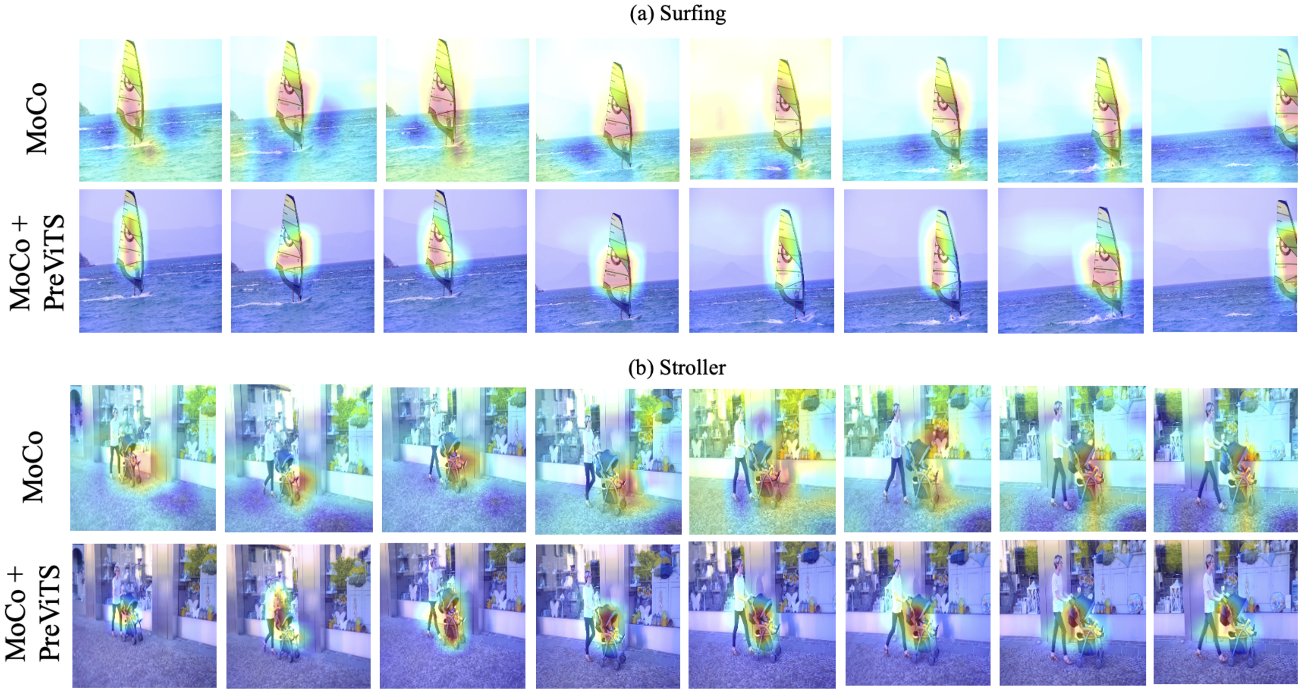
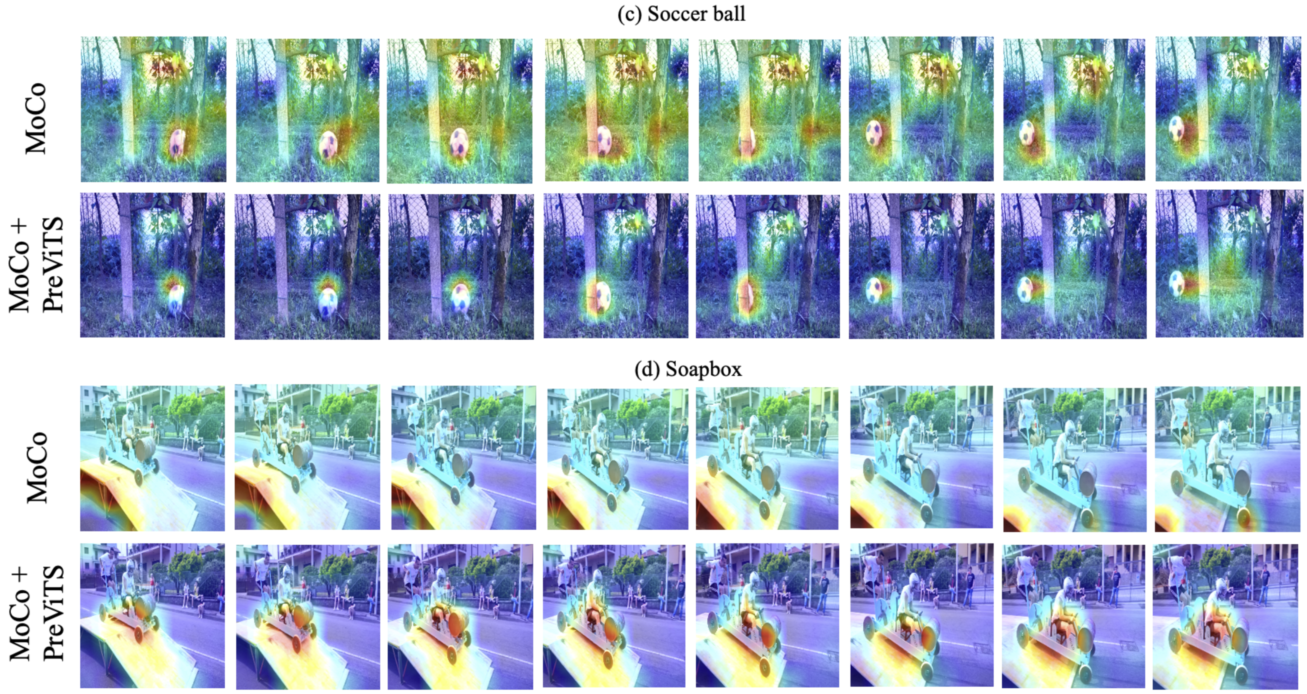


Figure 4: Grad-CAM Visualization for Video Backgrounds Challenge.


(a) Surfing

(b) Stroller

Figure 5: Grad-CAM Visualization for DAVIS Video Object Tracking and Segmentation.


(c) Soccer ball

(d) Soapbox

Figure 6: Grad-CAM Visualization for DAVIS Video Object Tracking and Segmentation.

References

[1] 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. 1

[2] 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. 1

[3] 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*, 2019. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009. 1

[5] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

[6] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2009. 1

[7] 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. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1

[9] 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. 1

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[11] 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. 1

[12] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[13] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[14] 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. 2

[15] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 1