We thank the reviewers for their feedback. We appreciate that all the reviewers agree on the value of our analysis and they like the intuitive idea of augmentation via videos. We also acknowledge that more experiments in 4.2 (R3) would help with justification/interpretation of results (R2). We are pleased to report that we have conducted several experiments that address *all* the concerns of the reviewers. Specifically, we perform two additional experiments to highlight how aggressive augmentation hurts (and it is not just the domain-gap). We have also conducted experiments to show that on a different downstream task, *video-based augmentation outperforms even ImageNet-based MoCo*. The goal of this paper is to present a critical analysis so that the community can also introspect as we make rapid progress on the topic of representation learning. We hope the discussion here and additional experiments will convince the reviewers and AC that this message deserves a wider audience at NeurIPS.

R2: Domain Gap or Augmentation effect? R3: Additional Experiment in 4.2 We agree: there is some ambiguity since the training data for COCO-cropped model is domain-aligned with PASCAL-cropped test data (unlike COCO-full). The drop in performance could be attributed to domain gap between COCO-full and PASCAL-cropped. So, we conducted a new experiment that cannot be explained by domain-gap.
Setup: Consider the subset of Pascal VOC07 images which depict either table or chair in the image, but not both (i.e scene-centric images containing only one of the frequently co-occurring pair of objects).
Result: On the table vs chair full image classification task, the representation trained on COCO-Boxes outperforms full COCO-image pre-training - 74.92 vs 73.64 mAP . Note that the test domain (PASCAL-full images) is aligned with training domain (COCO-full images). Yet, COCO-Boxes model outperforms COCO-full. This indicates that the problem lies with aligned representations of co-occurring objects (e.g., chairs and tables) as explained in the paper.

To support it even further, we do another experiment. We compare supervised learning on ImageNet with MOCO on ImageNet but for the task of PASCAL-part classification. Note when training on ImageNet, MOCO's cropping augmentations will learn embeddings that have similar representation for different parts. Hence, our augmentation interpretation predicts MOCO should not be as robust as supervised models on part-classifcation task. And indeed! the results indicate MOCO performance falls dramatically on part-classification.

Method	Dog Parts	Cat Parts
ImageNet R50	69.14	71.31
ImageNet MOCOv2	51.91	56.26
(Ours) Region Tracker	54.70	58.81

PIRL/MOCO should have higher Illumination Color Invariance (R2) The natural illumination color changes for measuring invariances are restricted to illumination ***temperature*** changes from 2175K to 3075K - this is also naturally captured in the ImageNet dataset. The synthetically color-jitterred samples are significantly different from these images depicting arbitrary color changes (see adjoining figure). On the other hand, occlusion is very easy to accurately synthesize by simply cropping images. And therefore the cropping augmentation strategy indeed leads to high occlusion invariance even on natural images.

Video-based Approach - Aggressive Augmentation and Performance (R1, R2) Our method only uses the aggressive augmentation in the Frame Temporal Invariance loss. The other component of the loss involving region tracks does not employ the aggressive cropping strategy. Including the aggressive augmentations ensures that we can take advantage of the occlusion invariance that it induces, while being constrained by the region-level loss. This extra constraint ensures the best of both worlds. This method was presented as a proof-of-concept to demonstrate that the appropriate invariances can be induced by using video data. Therefore, we did not scale up the data for this approach up to be comparable to ImageNet in terms of dataset size. However, we are pleased to report (as **R2** points out) that on a downstream task which is ill-suited to ImageNet-MOCO (due to aggressive augmentation), our video approach outperforms it (even with order of magnitude less data). The results on VOC-part classification shows that trend (See table above), also demonstrating that our approach is not affected by the aggressive augmentation.

RIS for Measuring Invariances (R3) Goodfellow et. al. [26] explains how the firing frequency can be used to measure invariances. We directly adopt the same principle. To the best of our knowledge, this is the most relevant metric proposed in past literature. It is true that two somewhat distinct representations after thresholding can have the same invariance under this metric. However, this is an intended feature proposed in [26] since most downstream classifiers would threshold a function of the neurons (generally linear classifiers and non-linear models in some cases).

Object Detectors as Objectness Prior (R1): We use a fully-unsupervised class-agnostic region proposal method - Selective Search. One should view this as a simple low-level preprocessing operation which ensures object-centric signals necessary for the augmentations. Since it uses no learning, we believe the comparisons are fair.
Argument on Capacity (R1): In light of new experiments, we agree with R1 and plan to remove this as it has no connection to main story. **Finetuning representations (R4)**: For this work, we wished to analyze the representations learned in the contrastive learning framework. Recent SSL benchmarks have suggested avoiding finetuning [25] since it would cause the representations to deviate from the initially learned representations, leading to spurious inferences.