Related Works. Thanks for the reviewers to point out the missing references in our initial content, which will be added in our final version. Paper "Video-Induced Visual Invariances" focuses on applying different pre-text tasks on frame/shot (rotation or frame level augmentation consistency) and video (future shot prediction consistency like DPC) respectively without reporting the performance on temporal related benchmarks. However, CCL focuses on using the nature of belong and inclusion cycle-consistency relation across the frame&video with contrastive representations in each domain by introducing a differentiable soft nearest neighbour formulation to train a single network e2e.

Method Details. As R1 pointed out, our backhone is not a typical 3D ResNet due to the motivation of maintaining the frame-level information before the temporal convolution for gathering video-level information described as L178-179. We will change the name of 3D ResNet in our model to "2D+1D ResNet". **As comments from R2**, we use soft nearest neighbour due to that nearest neighbour searching in CCL is not a differentiable procedure. Soft formulation for nearest-neighbours refer to [Neighbourhood Components Analysis] is a method by introducing a differentiable cost function based on stochastic neighbour assignments, and we introduce this idea into our approach to enable the cycle-contrast procedure to be e2e learnable. **As R3 pointed out**, the definition of $dis()$ is a cosine similarity shown as L134, and we will change the name of $dis()$ to $sim()$ in our final version. Our implementation is to avoid the query sample measuring the similarity with itself in the denominator, and we confirmed that indicator function used in Eq.(2),(4) is not an appropriate way and will be removed in our final version. The index of z will be modified according to **the advice of R2**, $n \in N$ in L104 will also be removed, and section 3.6 will be moved to Related Works. Regarding penalisation term P, we show an ablation on it in Table.6, and we acquired a further improvement by adding P with 1.8 points on UCF. **Further comments from R4**, we will add a detail description in the caption of Figure 1 with the architecture information of each component and the math symbol used in the Figure to make it self-contained. **As R1,R2 pointed out**, the balance parameter of Eq.(6) are 0.4, 0.4 and 0.2 for L_f, L_v and P respectively for both retrieval and action recognition settings. They will be added in the final version. We will also add a plot of our training and validation curves for each loss in our final version. **As R1 and R3 suggested**, the term "Transformer" in Figure 1 will be replaced by "Feature Projector".

Evaluation Details. As comments of R1, 8 frames are sparsely sampled from M-frame length video at a temporal stride as $\frac{M}{8}$ as described in L90-91 and L176-177. We will further clarify it in the caption of Table 1. The frame chosen from test clips is based on the same way as the above. For UCF, HMDB51 and Kinetics, video fps are 25 and for MMAct is 30. **As R1 and R2 pointed out**, TCC is not compared in our experiment. It is because we focus on the downstream tasks under a fully unsupervised setting. TCC still needs prior knowledge about whether the clips are the same action class or not. However, we agree with that it is interesting to check the results on other tasks, such as phase classification, to further prove the effectiveness of CCL by comparing with TCC which focuses on only frame level relation. We will add this experiment in our final version. **Further comments from R2**, similar as the settings in TCC for checking the cycle-consistency effectiveness, in Table 2, we use MSE as our baseline which is a straightforward way to make the frame and video satisfy the belong and inclusion relations. To further clarify the effectiveness of CCL and avoid the confusion regarding Table 2, we will remove Table 2 and merge the MSE result into Table 3 and add a result of MSE in loss ablation study in Table 6. The reason why we didn't compare with CBT and AVSlowFast in Table 3 and 4 is because these two methods are utilizing multi-modal information (CBT:video+language, AVSlowFast:video+audio) for self-training, which is not a fair comparison with the other methods listed in Table 3 and 4 that only use the video modality. However, **as R2 pointed out**, to give a complete picture of this topic, we will add the results of these two methods in table 4 as reference in our final version. DPC is a state-of-the-art approach by utilizing the idea of instance discrimination on frame level. In fact, we reported the result about the comparison with it in Table 4 and outperformed it with +1.2 points on UCF and +3.3 points on HMDB51 under the similar setting. As the res block in CCL are 2D conv, we achieved the above results with only **12.1M #param** that is lower than DPC with 14.2M. It is further confirmed that the effectiveness of CCL. **As R3 pointed out**, the retrieval experiment follows the reported results as paper COP(2019) and SpeedNet(2020). For further fair comparison, we will add the number of parameters of each model in Table 3 and 4. The reason why we use the clip from the test set to query the clips in the training set is because we should follow the same experiment setting as previous works' in Table 3. Moreover, using unseen data to query the seen data (initialized database for querying) is also a real use case in the retrieval system such as EC service. **As R4 pointed out**, we reported the e2e fine-tune result due to the fair comparison with the previous works that mostly done under this setting in Table 4. Result about fixing the backbone and fine-tune the FC layers was reported only in Table 5 as CCL(FC). We evaluate our approach under linear classification protocol on UCF and HMDB51 with 52.1 and 27.8 acc. respectively. It is higher than ShuffleLearn(58.3M #param) and 3D-RotNet(33.6M #param) with +25.6 and +4.4 points respectively, and slightly inferior to CBT that using multi-modal and S3D without providing the #param information according to CBT paper. We did not refer to paper SpeedNet as it was released after our initial submission, which will be added in the final version. SpeedNet achieves 81.1 by using a larger two-stream backbone S3D-G with 64 frames as input, while ours is only 8 frames with RGB stream. They also report the result using a weaker backbone I3D (25M #param according to paper [A New Model and the Kinetics Dataset]), showing a result about 66.7 which is lower than ours, and performed worse than all metrics in retrieval task in Table 3 with S3D-G. We will add all the above results to our final submission.