We thank the reviewers for their valuable comments.

R1, R2, R3, R4: Summary and novelty: Recent works have shown that the accuracy gap between supervised and self-supervised (SSL) models can be dramatically reduced by increasing the model capacity. We propose a simple yet effective unsupervised compression method for SSL representations that reduces the gap in less complex models. Our experiments compare with other compression methods and also show that using our method along with a strong SSL model outperforms all SOTA SSL methods. All reviewers agree that our method is simple, effective, and rigorously evaluated against competent baselines on different datasets and tasks. Below, we address concerns from the reviewers.

R1, R2, R4: Table 2: We agree that highlighting in this table is confusing. For a fair comparison, we did try CC with R50x4 teacher. The results are shown in the table beside. Our model outperforms CC when the teacher is R50x4 and is on par with CC when the teacher is R50. We will clarify this in the text.

AlexNet		CUB200				Cars196			
Method	Teacher	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
CC	R50	**23.2**	32.5	**45.1**	**58.2**	**23.7**	31.4	41.1	52.4
Ours	R50	23.1	**33.0**	**45.1**	58.0	23.6	**32.8**	**42.9**	**54.9**
CC	R50x4	23.6	33.6	44.9	58.4	25.4	33.2	43.2	54.3
Ours	R50x4	**26.5**	**37.0**	**49.4**	**62.4**	**28.4**	**38.5**	**48.7**	**60.4**

R1: Results of Section 4.4: Our goal is to show that our distillation method can compress SSL representations. In Sec 4.3, we show that our method is better than previous compression methods. In Sec 4.4, we aim to quantify its improvement over SOTA SSL models. We do not claim a stand-alone SSL method. Our method can improve upon other SSL methods when they are used as teachers. **Tables 3-6:** We will improve the organization of those tables. **FLOPS:** We will add FLOPS to the final paper. Compressing ResNet50x4 to AlexNet reduces computation by almost 80 times.

R2: Ablation: Yes, one can use the teacher's embedding for student's anchor points, but our goal was to decouple the embedding spaces of the teacher and student so that they can easily use different architectures. But, it's a good idea and we will add this experiment. Note that we show the effect of changing the queue size in Fig. 3.a. **BN layer:** Yes, we simply whiten each dimension by the batch statistics using the implementation of BN module. We will clarify. **Momentum:** Yes, we agree. Indeed, we show in Line 243 that momentum does not affect our results much.

R3: "Compressing unsupervised model now is meanless, as long as unsupervised methods are still evolving dramatically": We respectfully disagree. All machine learning algorithms are evolving and we should not wait to develop something novel. We believe there is some misunderstanding that our method is specifically designed for contrastive learning (CL). This is not true. Our formulation looks similar to CL, but we never compare representations of the same image. Instead of having positive or negative pairs, we compare each image with all "other" images, called anchor points, and transfer the actual similarity values.

R3: Surpassing supervised AlexNet: As pointed out in Lines 252-255, we are not claiming that our SSL method works better than supervised models "in general". We simply compare with the supervised AlexNet that is trained with cross-entropy loss, which is standard in the SSL literature. One can use a more advanced supervised training e.g., compress supervised ResNet50x4 to AlexNet, to get much better performance for the supervised model. We will clarify this more in the paper.

R3: COCO and ResNet50: Since our main focus was on compressing to smaller models, we used AlexNet for transfer experiments, which is standard in self-supervised learning community. Note that transferring to COCO needs more resources, and we have already done lots of experiments as acknowledged by other reviewers.

R3: NN is not always faster than linear classifier: In Lines 42-44, by "evaluation", we meant evaluating the SSL features (training a linear model and testing it). NN is faster since it does not need any training and parameter tuning. We will clarify in the final version. **Removing cross entropy loss from supervised distillation methods:** Yes, we agree. This is exactly what we did when using CRD (SOTA method) in our experiments. Note that CRD [43] is cited on Line 79. **Related work:** Thanks for pointing out those two papers. We missed them as they are not closely related and are not peer-reviewed. We will add them to the final paper. **Citing [a]:** [a] is a concurrent work that appeared on ArXiv just 17 days before our submission. We will cite it. **Self-distillation:** Our goal was to compress SSL models, so we did not consider self-distillation. That can be an interesting future work.

R3: Table 1 is a bit messy: We will improve its organization. We chose number of clusters for CC from the analysis in ClusterFit [49] and tuned the other parameters (learning rate, its schedule and number of iterations). CRD [43] already has a non-linear projection head and the paper shows that it is not very sensitive to the number of negatives beyond 4,096. We do not know why MobileNet-V2 is different from AlexNet in comparing CC with CRD.

R4: Why AlexNet is better: Yes, we agree that our method does not improve over supervised version of other architectures, but it consistently reduces the gap between supervised and unsupervised models. In general, the deeper the teacher compared to the student, the more improvement from compression methods. In out experiments, this ratio is maximized when we compress ResNet50x4 to AlexNet. We are using standard optimization for all architectures but AlexNet does not benefit from recent architectural tricks like BN and Residual layers.