General considerations. We thank the reviewers for their detailed comments and suggestions. We are glad that the response has been mostly positive with the reviewers finding the method interesting and the empirical investigation extensive. Before responding to each reviewer individually, we provide general considerations about experiments on ImageNet with larger backbones, which has been identified as an issue by some of the reviewers. We will attempt to get ResNet50/ImageNet runs for the final version. Recent self-supervised learning methods are expensive: SimCLR (ResNet50, ImageNet) has been trained on 128-v3 TPUs cores, using a very large mini-batch of 4096 (67.1×10^6 pairs). A fair comparison of our method against this setup would require a mini-batch of 1024 and $K = 256$ augmentations (66.8×10^6 pairs). To give an idea of the amount of computing required to run this experiment, consider that common GeForce RTX 2080 or Titan-X are close to saturation on CIFAR-10 with ResNet32, mini-batch 64, $K = 32$ and 6.3×10^4 pairs (two order of magnitude less than the desired target). To bypass this obstacle we provided extensive experiments on ImageNet surrogates such as tiny-ImageNet (200 classes, 120K images) and SlimageNet (1000 classes, 200K images). We guaranteed a fair comparison with SimCLR by equalizing the number of pairs (see Section 5 of the paper). Most importantly, we did not finetune our method; we used standard values for learning rates, focal loss, and weight decay, showing that unlike other recent methods we do not need heavy-finetuning. We consider this setup a reasonable compromise, as it prunes away inflated results and the need of large computational resources. Having said that, we will try to get ImageNet runs by allocating more resources and engineering a highly-parallelizable solution. However, it is unlikely that those resources will be enough for an extensive hyperparmeter search as for SimCLR. We also stress the fact that the code will be released, giving the opportunity to compare it on larger benchmarks.

Reviewer #1 marked the paper for clear acceptance (top 50% accepted papers), highlighting the efficiency of the method, and the thorough empirical evaluation in terms of backbones, datasets, and downstream tasks. We thank the reviewer for the positive feedback and the commentary, which has captured the essential strengths of this work. The reviewer pointed out that the paper can further its impact by evaluating on different tasks and larger benchmarks, stating that those experiments are not expected for the rebuttal. We have focused on classification and image retrieval, but we agree on the importance of extending the evaluation to other tasks, this will be likely done in future work.

Reviewer #2 recommends acceptance as slightly above threshold, highlighting the strong results, the clarity of the method, the extensive empirical investigation and the fair experimental procedure. The reviewer pointed out that it may be necessary to test the methods with larger backbones since self-supervised models may vary in performances with different backbones. We agree with the reviewer, as explained above we did not run those experiments for computational barriers. We will attempt a ResNet50 experiment on ImageNet for the final version.

Reviewer #3 marked the paper as slightly below threshold but found the idea interesting, in particular the fact that a binary cross-entropy loss can lead to strong results. We thank the reviewer for providing such a detailed commentary and hope that the considerations below will resonate with the reviewer. ***(1) Training procedure.*** The reviewer states that the procedure of training and testing on the same dataset is a popular but biased way of evaluating self-supervised methods. As the reviewer correctly pointed out, this procedure (also known as "linear probes") has been widely adopted, being used in a number of highly influential papers. This should be considered as a criticism to the way the community is benchmarking methods rather than a criticism against our setup. We need (as a community) to define and validate standardized protocols for self-supervised learning. Given the lack of such protocols, we have used what is common choice to make our work comparable. ***(2) Transfer learning experiments.*** There is scarce agreement on appropriate datasets and we consider this a corollary of the previous point: a general standardization issue. CIFAR 10/100 are a good compromise since they have same image size, they are popular, and provide a realistic distribution shift. Models trained on CIFAR10 must generalize to 90 unseen classes on CIFAR100 some of them remarkably different from the base set (e.g. furniture, electrical devices, etc), our method excels in this setting. Additionally, note that we have achieved SoTA with ResNet34 on STL-10 (see Table 1, appendix B.4), which has a domain shift between unlabeled and labeled set. ***(3) Clarifications.*** Focal loss: there is a minor but consistent improvement from the use of focal loss ($+0.5\%$ on average). Evaluation: we reported the accuracy after a fixed amount of epochs (test set has never been accessed during training). We also tried model selection through validation (held-out from training set); this did not change the difference in accuracy between methods. ***(4) Ablations.*** The analysis of the reviewer about the three main differences with SimCLR is correct. We will try those ablations in future work and add a clarification in the paper.

Reviewer #4 recommends acceptance as above threshold, highlighting how the method is an interesting alternative to contrastive approaches. ***(1) Difference in performance due to the loss.*** This was a speculative claim from our part, we agree that it may not be the only factor at play. We can make it clear in the final version. ***(2) Cross-entropy has been overlooked.*** We meant overlooked in the context of contrastive methods. NCE is a type of cross-entropy that requires the estimation of an expensive partition function; this has been a default choice in SoTA methods (this is why: "overlooked"). In our case, the use of a standard binary cross-entropy overcomes this issue. We can clarify this point for the final version. ***(3) Comparison with MoCo v1/v2.*** We gave higher priority to SimCLR rather than MoCo-v1 (same family but the former is stronger). MoCo-v2 was released in mid-March, we did not have time to implement it in our framework (for the NeurIPS rules a comparison was not due).