We thank the reviewers for their positive comments and detailed feedback. Specific concerns are addressed below.

R4: "Motivation unclear; compare to earlier metric learning work": In our related work (lines 97-109), we outline the differences between modern contrastive learning methods (including ours) and the papers listed by R4. In summary: our approach can be seen as a generalization of the prior approaches by using multiple positives and negatives per anchor; the use of a temperature parameter which plays the role of a margin; and the use of multiple views of the data. Each of these components contributes to the performance of the supervised contrastive loss, as shown in our ablation experiments. It is well known that distance metric learning approaches such as triplet loss and the loss proposed in [4] have slow convergence and require proper tuning of hard negatives (e.g. see [1], Section 3). Additionally, some of these losses seem to require very small batch sizes to avoid falling into local minima (e.g., [3] cites a batch size of 16 for this exact reason). Our approach overcomes these issues as shown by our application to large scale problems such as ImageNet. In Section 3.2.3, we show analytically the reduction to the triplet loss in Section 3.2.3. We will further clarify these points in the final version.

R4: "Comparison to metric learning methods": We tested N-pairs [2] in our framework with a batch size of 6144. N-pairs achieves only 57.4% Top-1 on ImageNet (compare to our loss that achieves 78.7%). We believe this is due to multiple factors missing from N-pairs loss compared to supervised contrastive: the use of multiple views; lower temperature (e.g., see [5], Table 5) and many more positives. We show some results of the impact of the number of positives per anchor in the Appendix (Sec. 6), and the N-pairs result is inline with them. We also note that the original N-pairs paper [2] has already shown the outperformance of N-pairs loss to triplet loss. We will add this experiment and relevant ablations to the paper.

R1/R2/R3: "Concerns over compute cost": We agree with the reviewers that reducing the expense of contrastive methods in general would be desirable. To this end, we experimented with memory based alternatives [1]. On ImageNet, with a memory size of 8192 (requiring only the storage of 128-dimensional vectors), a batch size of 256, and SGD optimizer, running on 8 Nvidia V100 GPUs, the supervised contrastive loss is able to achieve 79% top-1 accuracy on ResNet-50 architecture. This is in fact slightly better than the 78.7% accuracy with 6144 batch size (and no memory); and with significantly reduced compute and memory footprint. We will add this experiment to the paper. Finally, the drop in performance by training for fewer epochs (e.g. 350 vs 700 epochs) is very small (difference of 0.1%). We will add a sweep over training epochs to the paper.

R3: "Experiment reporting is confusing": We agree with R3 and will adjust the table reporting to be more clear and will highlight further SimCLR's performance for transfer learning.

R1/R3: "Improved baselines for cross-entropy; additional experiments": R3 requested that we show comparisons of cross-entropy and supervised contrastive losses using the same optimizers. In Table 4 of the Supplementary we include a sweep over optimizers used for supervised contrastive. In Table 1, we provide an equivalent sweep over optimizers used for cross-entropy using the same network, augmentation strategy, and number of train epochs; cross-entropy performs best with the momentum optimizer (the results show no improvement over those reported in the paper). We additionally ran experiments to see if the larger effective batch size used for supervised contrastive could explain its superior performance, as R1 suggested. We trained with cross-entropy loss using a batch size of 12,288, but this only achieved 77.5% Top-1 accuracy. We also tried training for twice as many epochs (1400) with the original batch size of 6,144, but this was even lower, at 77.0%. We will add these results to the final version. Please note in Figure 1 and Table 3 of the paper, baseline numbers are taken directly from previous papers.

R2: "Table 4 cross-entropy outperformed Supervised Contrastive; consider a joint loss": Note that Table 4 is about transfer learning where we see that SimCLR (unsupervised contrastive learning) actually outperforms both cross-entropy and supervised contrastive slightly. We agree with R2 that a joint cross-entropy and supervised contrastive loss could probably combine the benefits of both and certainly make it simpler to use the supervised contrastive loss in practice. We have experimented with this and seen strong results, although none that so far outperform the two stage approach. We will add a discussion of this in the final paper, but we feel that it's important for the sake of future research to isolate the impact of the supervised contrastive loss, so we chose to make the two-stage approach the focus of the paper.

Optimizer	Top-1 Accuracy
LARS	76.2%
RMSProp	76.2%
Momentum	78.2%

Table 1: Cross-entropy (ResNet-50) with different optimizers.

References

[1] Momentum Contrast for Unsupervised Visual Representation Learning, He et. al., 2019.
[2] Improved Deep Metric Learning with Multi-class N-pair Loss Objective, K. Sohn, 2016.
[3] Discriminative Feature Representation for Person Re-identification by Batch-contrastive Loss, Zhang et. al., 2018.
[4] Dimensionality reduction by learning an invariant mapping, Hadsell et. al., 2006.
[5] A Simple Framework for Contrastive Learning of Visual Representations, Chen et. al., 2020.