We thank each of the reviewers for their time to review the paper and for providing feedback on the submission. We provide a brief summary of the reviews, and detail each reviewers concerns regarding the submission individually: The submission is well-written [R1] ,[R2] ,[R3] and proposes a simple and effective idea [R1] , [R2] , [R3] which can be applied in a diverse setting to improve learning in CNNs [R1] , [R3] . The idea is well motivated [R3] and elegant [R2] , and outperforms the baselines in diverse contexts [R1] , [R3] such as image classification, feature learning, generative models, among others [R1] , [R3] , [R4] . The experimental results are convincing across the different settings [R1] , [R3] . The code submitted is simple, and easy to follow [R1] , [R3] which aids reproducibility [R1] , [R2] , [R3] , [R4] .

[R1] : The choice of using a single kernel size for all the layers of the network. We provide a discussion over the effect of adding CBS to single layer of the network in Appendix A, but one of the main benefits of using CBS is that the method works well without extensive hyperparameter tuning. It may be possible that the empirical performance can be further improved by adjusting the kernel-size and for the different layers in the network, but simply using a kernel-size of 3 strongly suggests the robustness of the proposed solution.

[R2] + [R1] : there exist many techniques beyond SGD to improve training (eg, as mentioned by the authors, batch normalization), against which the authors do not benchmark. The main benefit of the paper is that **CBS augments the already existent CNN improvements, such as batch normalization and dropout**. More specifically, techniques like batch normalization already exist in ResNet, WideResNet and ResNeXt architectures; similarly VGG network utilize Dropout. Since each of the network architectures are used as proposed, **this strengthens the empirical evaluation since CBS improves models without changing the underlying CNN improvements that already exist**.

[R2] : choice of σ and its decrease rate are not theoretically studied: We provide a discussion over the choice of σ and the decay rate in the Appendix C, empirically. We will elaborate on the choices in the paper for the final draft.

[R3] : analyzing other smoothing and compare the results to understand the choice of the Gaussian kernel. We analyze the effect of different smoothing curricula in Table 7. The different common kernels that exist are linear kernels, box-kernels (similar to average-pooling), Laplacian kernel, among others, which do not provide smoothing. There may be kernels of different forms that have similar behaviour in practice, but a Gaussian kernel is effective and easy to implement, and has been studied thoroughly in signal and image processing.

[R3] : Does the smoothing accelerate the optimization? We perform additional experiments and monitor the epoch number when the best validation accuracy is reached for WideResNet trained on CIFAR-100. For both the baseline and CBS the best validation accuracy is reached at epoch 90.8 ± 0.6 and 91.0 ± 0.8, which suggests that **the optimization is similar to the baseline, while achieving better performance.** Similar results are seen for other architectures.

[R4] : there are plenty of curriculum learning methods that could have been used as relevant state-of-the-art competing methods to compare with. Curriculum learning papers work by adjusting the order of sampling or using importance weights, which is orthogonal to CBS, since the batch sampling modifications can be added on top of CBS as well. But for completeness and given the short rebuttal period, we run the comparison against [R2] in the suggested references, since its closely related to [R3] as well on the CIFAR-10 dataset using importance weighting based on image difficulty, as proposed in [R2]. **For both experiments the suggested baseline does not improve vanilla CNNs**: 91.7 ± 0.3 and 92.7 ± 0.2, for Wide-ResNet and ResNeXt, respectively. One possible reason for why weighting-based methods do not outperform a vanilla CNN, in supervised learning, may be due to overparameterization of NNs resulting in memorization of all data as suggested in [1] [1]. CBS does not depend on importance weights or data-sampling. CBS is also inherently simpler than typical curriculum learning methods, as it does not rely on a measure of "task-difficulty" compared to the citations suggested by Reviewer 4 [R2, R3]. We will add the results and discussion to the final draft.

[R4] : non-linearity is typically applied before pooling We will fix the equation as suggested in the final draft.

[R4] : Does the approach apply to data other than images? We study the effect of training CNNs with a Gaussian filter since images, unlike text or tabular data, have strong relations to signal processing. The inherent simplicity of the technique make it possible such that similar benefits can be gained on non-image data, but typically it is sufficient for a paper to show experiments with only data of one modality. The diverse nature of experiments does suggest that CBS improves CNNs in a variety of different contexts beyond supervised learning.

[R4] : Are the improvements statistically significant? In machine learning literature, the standard of reporting the mean and standard deviation with 5 random seeds is performed for all experiments. We do note that for most experiments, the sum of the baseline and its standard deviation is lower than the score for CBS minus its standard deviation, which suggests that the results are indeed statistically significant.

[R1] [R2] [R3] [R4] Missing citations and typos Thank you for the suggestions; we have made the suggested corrections.

[1] [1] Byrd, Jonathon, and Zachary Lipton. "What is the effect of importance weighting in deep learning?." ICML. 2019.