**Technical novelty:** we acknowledge R#1 and R#2's concern that our method is technically relatively simple. However, despite the simplicity of our technique, we achieve performance that is **state of the art (by a large margin) for certified robustness to patch adversarial attacks on CIFAR-10**, as well as **the first scheme for certified robustness to patch adversarial attacks that scales to ImageNet.** We believe that, due to this improved performance, our algorithm warrants consideration: as R#3 states, the simplicity of our algorithm "is perhaps a plus rather than a minus." These significant empirical gains are results of our contributions in this paper including (i) the proposal of the structured ablation methods, and (ii) the proposal of de-randomization in patch smoothing.

**Inference/certification time (R#2):** In our column smoothing method, which empirically performs best, the number of forward passes required equals the width in pixels of the image (28 for MNIST, 32 for CIFAR, 299 for ImageNet). This improves from randomized smoothing methods, which typically use $\geq$ 10,000 forward passes per image. Empirically, with one GPU, we certify the entire 10,000 image CIFAR test set in less than 42 seconds, and the entire MNIST training set in less than 11 seconds. On 4 GPUs, certifying ImageNet images averages less than one second per image, amortized. We will add exact times to the revised draft.

**Empirical Robustness to Black-Box attacks:** R#2 asks for evaluation against empirical gradient-free patch attacks. Upon the reviewer's suggestion, we used the ARMORY Adversarial Robustness Evaluation Test Bed, and tested using the Universal Patch scenario. In this scenario, the attacks are generated independently of the attacked model (blackbox attack). The dataset tested is RESISC-45 (aerial photograph classification, with 45 classes and $224 \times 224$ image resolution: in other words, the image scale is similar to ImageNet, but the number of classes is considerably fewer). The attacked patches vary in size, and cover at most 25% of the area of the image. Using the hyperparameters borrowed from ImageNet from the paper (column smoothing, $s = 25, \theta = 0.2$), our model achieved 66% accuracy under attack and 80% clean accuracy, versus 23% accuracy under attack and 93% clean accuracy for the reference, undefended classifer provided for the scenario. This shows that our model substantially reduces the effectiveness of the attack, even with no hyperparameter optimization. Our model also had 52% *certified* accuracy against $42 \times 42$ adversarial patches. We will include these results, as well as experiments on non-patch adversarial distortions.

**Dependence on block size (R#1)**: The block size (s) in our defense controls the amount of information the base classifier can use, and so is directly analogous to (the inverse of) the smoothing standard deviation $\sigma$ in standard Gaussian randomized smoothing (Cohen et al. 2019). Like in that work (and indeed in all randomized-smoothing techniques that we are aware of), there is a tradeoff between robustness and accuracy, with accuracy falling at high variance (low $s$), and robustness falling at low variance (high $s$). The fact that $s$ must be the same at training and test time is also true of $\sigma$ in standard randomized smoothing.

**Organization Suggestions:** We will take the reviewers' suggestions, to move content out of the paper's introduction, and to incorporate content from Appendix E into the main text to better highlight the effect of derandomization. We will also improve the presentation/readability of the data in Figure 5, and remove the redundant presentation of information in Tables 1, 2 and Figure 1.

**Role of $\theta$ parameter (R#2):** As we mention, the hyperparameter $\theta$ and the thresholding scheme only have a significant effect on MNIST, not on the more realistic CIFAR-10 dataset. We will explain this more clearly in the revised draft.

**Ethical aspects of robustness (R#2)** We will add further discussions about ethical concerns in robust ML to the paper.

**Low accuracy on ImageNet (R#4):** We acknowledge that our clean accuracy on ImageNet is relatively low (44.6%). However, given that ours is the *first* work to report patch-attack certificates at ImageNet scale, we still believe that our result on ImageNet has merit. Also, we remind the reviewers that ImageNet is a 1000-class classification problem, so 44.6% top-1 accuracy (and 13.9% certified accuracy) are still highly nontrivial. For further results on a dataset with complexity intermediate between CIFAR-10 and ImageNet, see results on RESISC-45 ("Empirical Robustness" above).

**Why does Column Smoothing outperform Block Smoothing? (R#4)** As shown in Figure 3, even when the risk of sampling the adversarial patch is higher for column smoothing and the number of retained pixels is lower, column smoothing still outperforms block smoothing because the accuracy of the base classifier is higher (despite using fewer pixels). We can speculate that this is because the base classifier has access to more varied parts of the image when sampling a column, as compared to sampling a block.

**Comparisons to Chiang et al. (2020) vs Levine and Feizi (2020) (R#4, R#1)** We chose to first highlight the comparison to Chiang et al. (2020), because Chiang et al. (2020) presents a specific certified defense against patch attacks. By contrast, Levine and Feizi (2020)'s defense against *sparse* adversarial attacks only incidentally provides certificates against patch attacks: these certificates are not competitive at all for the patch threat model. The reason that we provide this comparison is because our *techniques* are related to Levine and Feizi (2020): it therefore makes more sense to provide the comparison when discussing our techniques (to show that structured ablation has a benefit over sparse randomized ablation), rather than as a "top-line" result.