	CIFAR10 2.4% patch			ImageNet 2% patch		
	ours	DRS	PG+DRS / Bagnets	ours	DRS smoothing	PG+DRS / Bagnets
attack success rate	83.2	94.3	92.1 / 87.3	78.5	84.9	83.2 / 79.3
classification accuracy	60.3	59.5	61.4 / 40.4	34.4	27.0	30.1 / 31.2

Table 1: Attack success rate for un-certified images and classification accuracy on all images (%) under sample-specific patch attack

1 reviewer 1

Thank you for the positive and helpful review. We have addressed the comments in the updated manuscript. Below please find responses for the questions:

While adversarial training is a strong empirical defense against adversarial attack, it does not provide robustness guarantees, i.e., given a clean/unperturbed image, an adversarially-trained classifier cannot certify if adversarial attack can or cannot change the predicted class on this image. On the other hand, our method can certify if the predicted class of a given image can or cannot be changed under patch attack. We have considered adversarially-training f_θ, and it does improve empirical performance but decreases certification probability. This is because certification probability increases as the number of crops from *clean* image being classified correctly increases, but adversarial training decreases performance on clean samples and consequently decreases certification probability.

A deterministic crop selection covering the whole image was considered, but we found that random selection is easier to select the number of crops, while having comparable certified accuracy.

We list the percentage of images that cannot be certified but can be successfully attacked by image-specific patches in the below table with the corresponding overall accuracy. We train our image-specific patch by training an adversarial patch which is placed at the center of the image for each image. We train each patch for 100 steps with step size 0.02. Please note that although some images are still correctly classified under image-specific patch attack, there can still exist a patch that can change its classification output. Due to time constraints, we only show CIFAR with patch size 2.4% and ImageNet with patch size 2% with no patch transformation.

2 reviewer 2

Thank you for the helpful review. We've added training time to the first paragraph of Section 4. Clean accuracy are shown in Figure 2 and the ablation study of p_c is added to Table 2.

We respectfully disagree with the reviewer that our method has only marginal improvement over PG+DRS – with no patch transformation, we achieve comparable certification accuracy as PG+DRS while the inference time is more than a magnitude faster; with patch transformation, our worst-case certification accuracy is significantly better than PG+DRS. This comes from our basic idea of using crops instead of ablating regions of an image.

3 reviewer 3

Thank you for the helpful review. We have made corrections to Eq. 6 and added experiments and discussion of confidence level in Appendix D.

4 reviewer 4

Thank you for the thorough review. We've added a brief overview of Derandomized smoothing and PatchGuard in Appendix A, pseudo code explaining how certification probability is computed in Appendix B, and derivations and experimental results for uniform sampling without replacement in Appendix C. We've also expanded discussion on how certification rate is reported at test time in the "metric" paragraph in Section 4 and the effect of crop size vs patch size in Section 4.4. We analyze all three components of f_θ in the "Training randomized cropping classifier" paragraph of Section 3.1 and show that f_θ has the same set of trainable parameters as g_θ. We had several passes of revising the manuscript to improve the readability.