To all reviewers who provided feedback, thank you. Based on the common comments about the writing, we will use your suggestions on how to improve on it. In particular, we will stop using the phrase "**breaking a backdoor defense**" to avoid confusion. Instead, we will conclude that "**existing backdoor defenses may not be applicable to an adversarially trained or robust network**." We would like to address your other concerns in the following:

To Reviewer 1. Q1: Does reducing the input gradients (Simon-Gabriel et al., 2019) opens up vulnerability to backdoor attacks? **A1:** Thanks for this good question. We implemented the work by adding a regularization term to the loss of a model. As we increase the strength of the regularization term, the adversarial robustness of the model increases as well on CIFAR-10 while the backdoor robustness decreases, as shown in Table 1. The input gradients give another way to explain the trade-off and we will add the results to our paper. **Q2:** Theoretically grounded experiments. **A2:** The certified robustness methods (see line 27) are theoretically grounded defense methods because they target the worst-case adversarial robustness. We showed in lines 151-157 that even a certified robustness method IBP (Sven Gowal et al., 2019) are subject to the trade-off. **Q3:** What are practical cases that need to consider both adversarial and backdoor threats against the same model? **A3:** Any security-sensitive model that may be under attack *during data collection* and *after training* should consider both threats. Examples span from self-driving cars that learn from public scenes to face-detection systems built on an open collection of face images. **Q4:** Unclear categorization of backdoor defenses. **A4:** Thanks. We will improve our writings based on your suggestions and avoid using the phrase "breaking a backdoor defense." We hope the above clarifies your concerns and convinces you to improve your rating.

Table 1	Input Gradient Regularization Strength			
	0	0.005	0.01	0.035
Adv. Robustness	0	0.007	0.025	0.132
Backdoor Succ. Rate	0.453	0.802	0.889	0.993

To Reviewer 2. Q1: The TRADES model may improve both the robustness and back-door robustness. **A1:** Thanks for this good question. We run TRADES on CIFAR-10 using the code and settings provided by the authors and found that the trade-off still holds, as shown in Table 2. **Q2:** It is unclear whether the trade-off still holds when the models that are partially adversarially robust. **A2:** This is an interesting direction to explore. We make a model "partially" adversarially robust by adversarially training it with a PGD attack that has a smaller ϵ (i.e., the maximum allowable perturbations to the input). This makes the model less robust when it is evaluated by a PGD attack with a larger ϵ at test time. Table 3 shows the results on CIFAR-10. As we can see, the trade-off still holds. In particular, the backdoor robustness of the model seems to degrade quickly as the adversarial robustness increases. **Q3:** Are the data for the adversarial training poisoned or not? **A3:** Yes. **Q4:** Would that mean successful backdoor attack also reduces adversarial robustness? **A4:** Possibly. But in practice, we observed very little difference in adversarial robustness across different models and datasets. **Q5:** Too few steps of attack for adversarial attack (only 5 to 10 steps), it may not access the true adversarial robustness. **A5:** Following your suggestion, we evaluate the adversarial robustness of the adversarially trained models using the PGD attack with 200 steps on MNIST and CIFAR-10. Table 4 shows the results, which indicate that the models have reasonable adversarial robustness. We hope the above clarifies your concerns and convinces you to improve your rating.

Table 2	TRADES	
	Reg. Trained	Adv. Trained
Adv. Robustness	0	0.543
Backdoor Succ. Rate	0.275	0.988

Table 3	Train-Time ϵ			
	4/255	8/255	12/255	16/255
Adv. Robust. ($\epsilon = 16/255$)	0.119	0.257	0.306	0.312
Backdoor Succ. Rate	0.993	1	0.999	0.999

Table 4	PGD Steps	
	5	200
Adv. Robustness on MNIST	0.93	0.92
Adv. Robustness on CIFAR10	0.45	0.39

To Reviewer 3. Q1: The main idea of the paper is quite interesting, and the intuition of the trade-off is quite clear in hindsight. **A1:** Thank you for your positive comments. **Q2:** Line 130: I interpret 5% of the dataset to correspond to 50% of the target class for CIFAR10, is this correct? **A2:** Yes. We also experimented with 10% and 5% of the target class in Section 4.2. **Q3:** How does this translate to ImagetNet? **A3:** 0.05% of training data means 50% of the target class. **Q4:** In spectral signatures, ... how is the fraction of removed training data determined for different poisoning ratios? **A4:** Here we assume the defender knows the number of poisoned examples, so we remove the same amount of examples from training data. This favors the defense. **Q5:** Does a "dirty-label" backdoor attack perform better or worse in the case of robust training than in standard training? **A5:** Good question. Our results showed that the attack achieves similar backdoor success rates in the two cases. We hope the above clarifies your concerns and humbly ask that if you think our findings deserve attention to the field, please champion this paper.