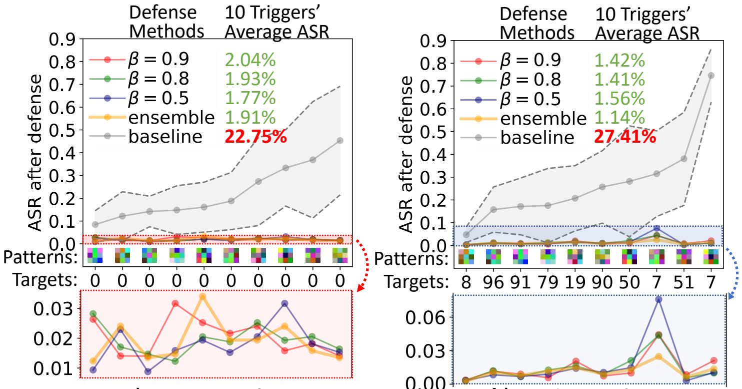
Clarification of the MESA method (Review #2, #4). We are aware that the dense notations in Section 3.1 is hard to follow. We will rewrite the part in a more descriptive language and move the excessive details to Appendix. For the question by Reviewer #2, g' is the derivative of g. The $g' > \omega$ in line 149 should be corrected as $g'(\beta) > \omega, \forall \beta$.



a) Results on Cifar 10

b) Results on Cifar 100

Figure 1: We generate 10 triggers with independent and uniform RGB colors, and test them on CIFAR-10/CIFAR-100 with target class 0/random.

The choice of dataset, target class, and black-white trigger (Review #2, #3, #4). Unlike adversarial attacks that are closely related to the dataset (especially the decision boundary), backdoor attacks inject arbitrary triggers that have little relationship with the dataset. The trigger is usually deliberately designed to be out-of-distribution to achieve a high attack success rate. From the feature space's perspective, we can assume that normal data from all classes form one cluster, and poisoned data form another cluster. Previous research [1] explicitly used this property for backdoor defense. The intuition is that the distance between normal data and poisoned data, instead of the dataset and class labels, determines the attack/defense difficulty.

Without worrying about the out-of-distribution issue, we chose black-white triggers for a better coverage of corner cases. Naively randomizing RGB colors with [0, 255] values can (almost) never generate special triggers such as a 3×3 black square (requiring 27 zeros). We add several experiments using random-color triggers as shown in Figure 1. Our defense method obtains $\sim$2% after-defense ASR (attack success rate) in average, better than the previous results on black-white triggers ($\sim$4%). The experiments are performed on both CIFAR-10 (Figure 1(a), fixed target class) and CIFAR-100 (Figure 1(b), random target class) to show the marginal effect of dataset and target class choices. In the final submission, we will include the discussions on the impact of trigger color, dataset, and target class.

Regarding to Reviewer #4's concern about the size of the support set, the choice of black-white and colorful triggers only decides the support set of original triggers, not reversed triggers. The search space of reverse engineering is always $\mathbb{R}^{27}$ regardless of the choice of original triggers.

Trigger locations (Review #4). We randomize the trigger location in each attacking step, and the backdoored model is sensitive to the trigger at any location. The defense algorithm tests a generated trigger at a random location in each step, and trains the generator with gradients from all locations. The only prior knowledge is the 3×3 trigger size.

Comparing to related works about model ensembling (Review #5). GWN [2] trains multiple models towards the same target distribution and introduces inter-model interaction to improve image diversity. DoPaNet [3] trains multiple discriminators targeting different modes to reduce the modeling complexity. Both methods use model ensembling as an enhancement to the original GAN and solve the mode dropping problem. Without model ensembling, a single GAN can still theoretically model any arbitrary distribution, if ignoring the model capacity limitation in practice.

The model ensembling in this work has a completely different motivation. In the sampling-free setting, the Nash equilibrium in GANs does not exist, and we have to train the generator without any discriminator. Under this constraint, we find that directly learning an arbitrary distribution being difficult, and simplify the problem by targeting uniform distributions. Then, model ensembling becomes a mandatory step to recover the arbitrary distribution from uniform distributions. A single model is incapable to capture the arbitrary distribution both theoretically and practically.

Advantage of model ensembling and justification of parameter selection (Review #5). The optimal backdoor defense is to retrain the backdoored model using the original trigger. Reverse engineering the trigger distribution is only an alternative approach to cover the original trigger. The ensemble model does not necessarily perform better than a single sub-model, since any sub-model has a chance to cover the original trigger.

Model ensembling mainly serves for two purposes: 1) use multiple cross-sections to allow us to better understand the shape of trigger distribution and 2) provide a robust defense without parameter tuning. Parameter α balances the hinge loss and regularization. Fixing α to any value between 0.1 and 1 is empirically okay. Parameter β is related to the strength of the attack and an inappropriate value leads to less effective defense (the reversed trigger distribution can be too sparse or too narrow to cover the original trigger). When the attack is unknown and β cannot be predetermined, an ensemble model provides a robust defense without tuning β (see yellow lines in Figure 1).

[1] Brandon Tran et al., Spectral Signatures in Backdoor Attacks, In NeurIPS, 2018.

[2] Honglun Zhang et al., Generative Warfare Nets: Ensemble via Adversaries and Collaborators, In IJCAI, 2018.

[3] Botos Csaba et al., Domain Partitioning Network, arXiv:1902.08134, 2019.