# EXPLORING VULNERABILITIES OF SEMI-SUPERVISED LEARNING TO SIMPLE BACKDOOR ATTACKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Semi-supervised learning methods can train high-accuracy machine learning models with a fraction of the labeled training samples required for traditional supervised learning. Such methods do not typically involve close review of the unlabeled training samples, making them tempting targets for data poisoning attacks. In this paper, we show that simple backdoor attacks on unlabeled samples in semi-supervised learning are surprisingly effective - achieving an average attack success rate as high as $96.9\%$. We identify unique characteristics of backdoor attacks against semi-supervised learning that can provide practitioners with a better understanding of the vulnerabilities of their models to backdoor attacks.

## 1 INTRODUCTION

In recent years, semi-supervised learning methods have significantly increased in effectiveness and gained in popularity. These methods train models with a small set of labeled data and a large set of unlabeled data while maintaining comparable classification accuracy to supervised learning. In this work, we examine the vulnerability of semi-supervised learning methods to backdoor data poisoning attacks in the unlabeled data which is unlikely to undergo detailed human review. Backdoor data poisoning attacks insert a backdoor into a trained model that can cause sample misclassification through the introduction of a trigger (Gu et al., 2017). Traditionally, this is accomplished by introducing triggers into poisoned images during training and adapting the images or the training labels to encourage the network to ignore the image content of poisoned images and focus on the trigger. Training labels play a critical role in attacks against supervised learning. Dirty label attacks change the training labels from the ground truth label (Gu et al., 2017) and clean label attacks maintain the ground truth label while perturbing the training sample (Turner et al., 2019; Saha et al., 2020; Zhao et al., 2020). In semi-supervised learning, backdoors must be introduced in the absence of training labels associated with the poisoned images. Instead they rely on pseudolabels estimated from model predictions, a strategy used in recent semi-supervised learning methods (Lee et al., 2013).

In this work, we analyze the impact of backdoor data poisoning attacks on semi-supervised learning methods by first reframing the attacks in a setting where pseudolabels are used in lieu of training labels, and then highlighting a vulnerability of these methods to simple and accessible attacks which influence expected pseudolabel outputs. We contrast the performance of clean label backdoor attacks on supervised learning to backdoor attacks against unlabeled samples in semi-supervised learning in order to highlight the need for practitioners to adapt their mindset when determining how susceptible their model or defense is to attacks. Additionally we identify characteristics of successful attacks and analyze unique dynamics of data poisoning during semi-supervised training.

Recent semi-supervised learning techniques that have significantly improved classification performance (Xie et al., 2020; Berthelot et al., 2020; Sohn et al., 2020) using the strategies of consistency regularization and pseudolabeling. Techniques that employ consistency regularization encourage similar network outputs for augmented inputs (Sajjadi et al., 2016; Miyato et al., 2018; Xie et al., 2020) and often use strong augmentations that significantly change the appearance of inputs. Pseudolabeling uses model predictions to estimate training labels for unlabeled samples.

Data poisoning in the context of semi-supervised learning is a relatively new topic area. Some work focuses on poisoning labeled data (Liu et al., 2019; Franci et al., 2022) or using instance-targeted data poisoning attacks (Carlini, 2021), both of which are out of the scope of this paper which focuses on backdoor attacks on unlabeled data. Yan et al. (2021) investigate perturbation-based attacks on

unlabeled samples in semi-supervised learning similar to us, but find a simple perturbation-based attack has low attack success, motivating their introduction of a more complex attack. By contrast, we show that simple perturbation-based attacks can be very successful in the right settings. Shejwalkar et al. (2022) has work concurrent to ours which also examines the vulnerability of semi-supervised learning to simple backdoor attacks in unlabeled data. They focus on defining the most effective trigger pattern for successful attacks whereas our work focuses on how modifications of the poisoned samples can influence pseudolabel behavior and vary the effectiveness of attacks.

## 2 Backdoor Attacks in the Context of Semi-Supervised Learning

### 2.1 Attack Threat Model

We consider a setting in which a user has limited labeled, trusted data and a large amount of unlabeled data which may be poisoned. They train their model using the FixMatch semi-supervised learning method (Sohn et al., 2020). The adversary introduces poisoned samples into the unlabeled dataset with the goal of creating a strong backdoor in the trained network and maintaining high classification accuracy. Because the poisoned samples are only included in the unlabeled portion of the training data, the adversary can only control the image content for the poisoned samples and not the training labels. The adversary does not have access to the user's network architecture.

### 2.2 FixMatch Details

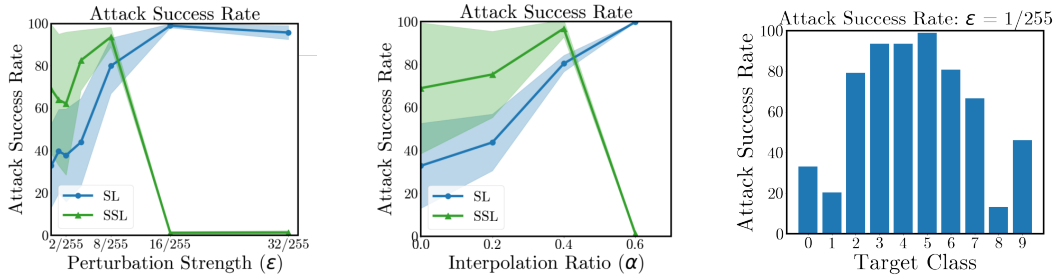
FixMatch achieves high classification accuracy with very few labeled samples using pseudolabeling and consistency regularization. FixMatch approximates supervised learning by estimating pseudolabels $\boldsymbol{y}^*$ for the unlabeled samples: $\boldsymbol{y}^* = \text{argmax}(f_\theta(T_w(\boldsymbol{u})))$, where $f_\theta(\cdot)$ is the network being trained and $T_w(\cdot)$ is a function that applies "weak" augmentations to the samples. If the confidence of the estimated label is above a user-specified threshold, the pseudolabel is retained and used for computing the unsupervised loss term. The unsupervised loss term is a consistency regularization term which encourages the network output of a strongly augmented sample to be the same as the pseudolabel estimated from the associated weakly augmented sample.

### 2.3 Backdoor Attack Vulnerability Considerations

With the consistency regularization and pseudolabeling in mind, we rethink how poisoned samples in backdoor attacks may interact differently in semi-supervised training than in supervised training. Prior work has shown that backdoor attacks are much less effective when data augmentation is used during training (Schwarzschild et al., 2021). Therefore, when poisoning samples in semi-supervised learning, it is important to use a trigger that is robust to both the weak and strong augmentations that are crucial to achieving high classification accuracy. Because of the reliance on pseudolabels, we suggest that attacks against semi-supervised learning be developed by considering how an adversary may vary the image content in a way that influences the expected pseudolabel outputs.

To analyze the impact of pseudolabel behavior on attack success, we use two types of attacks. The first attack, inspired by clean label backdoor attacks in supervised learning Turner et al. (2019), uses untargeted adversarial perturbations to influence estimated network outputs. The second attack uses poisoned images that are interpolated between target class samples and non-target class samples. Each poisoned sample $\boldsymbol{u_i^*}$ is defined as $\boldsymbol{u_i^*} = (1-\alpha)\,\boldsymbol{u_i^t} + \alpha \boldsymbol{u_i^n}$, where $\boldsymbol{u_i^t}$ is a sample from the target class, $\boldsymbol{u_i^n}$ is a sample from a non-target class, and $\alpha \in [0, 1]$ defines the interpolation ratio. In both attack types, by varying the strength of the modification, whether it be perturbation or interpolation, we can make target class samples more or less likely to result in target class pseudolabels.

To perform perturbation-based attacks, we use Projected Gradient Descent (PGD) adversarial perturbations (Madry et al., 2018), varying the constraint $\epsilon$ on the $\ell_\infty$ norm of the perturbation magnitude. For interpolation-based attacks, we interpolate between our selected target class samples and randomly selected non-target class samples while varying $\alpha$. In both cases, as the poisoned samples deviate more greatly from the original samples (by increasing $\epsilon$ or $\alpha$), fewer poisoned samples are estimated to be the ground truth label and the entropy of the distribution of network outputs increases, indicating the class estimates are distributed more evenly across all class outputs. Appendix A contains the detailed analysis of the impact of varying modification strength on the classification outputs for




(a) (b) (c)

Figure 1: (a) ASR for perturbation-based attacks against supervised learning (blue circle line) and semi-supervised learning (green triangle line) with varying $\epsilon$. (b) ASR for interpolation-based attacks. (c) ASR from a weak perturbation attack ($\epsilon = 1/255$) as the target class is varied.

CIFAR-10. Therefore, at low modification strength, we expect most poisoned samples have their ground truth classes as pseudolabels. At greater modification strength, we expect most poisoned samples will not have their ground truth classes as pseudolabels and instead their pseudolabels will be relatively evenly distributed across other classes.

## 3 Analysis

We analyze the vulnerability of semi-supervised learning methods to our pseudolabel-influencing attacks by considering the following experimental setup. We generate attacks using the CIFAR-10 dataset (Krizhevsky et al., 2009) and perform our analysis on FixMatch (Sohn et al., 2020). We largely follow the experimental details from (Sohn et al., 2020), using a WideResNet-28-2 (Zagoruyko and Komodakis, 2016) architecture, RandAugment (Cubuk et al., 2020) for strong augmentation, and horizontal flipping and cropping for weak augmentation. We experiment with 250 labeled samples. We limit each experiment to 100,000 training steps, finding that these shorter training runs achieve relatively high classification accuracy (around $90\%$) and attacks often reach a stable state long before the end of the runs. See Appendix B for more experimental details. We define the target class of the attack as the ground truth class from which we select samples to be poisoned. We modify the images then add augmentation-robust four-corner triggers. We analyze test accuracy and attack success rate (ASR) for determining the success of backdoor attacks against semi-supervised learning methods.

### 3.1 Success of Simple Pseudolabel-Influencing Attacks

Fig. 1 shows the results of our experiments investigating the impact of modification strength on attack success. Fig. 1a shows the performance of perturbation-based attacks as we vary $\epsilon$. For each $\epsilon$, we run five trials, varying the target class for each run from classes 0-4, and poison $1\%$ of the entire dataset. We compare the performance of the attacks against supervised learning (blue line) and semi-supervised learning (green line). The attacks against semi-supervised learning are highly successful for moderate perturbation strengths with an average ASR of $93.6\%$ for the attacks with $\epsilon = 8/255$ compared to an average ASR of $82.58\%$ for the attacks on supervised learning. There is a large variation in the ASR for attacks with weak perturbations and no perturbations. Fig. 1c shows the variation of ASR for a weak perturbation attack ($\epsilon = 1/255$) against semi-supervised learning as we vary target classes. While the ASR against supervised learning continues to increase with larger perturbations, the attacks fail against semi-supervised learning. Fig. 1b shows the performance of the interpolation-based attacks as we vary $\alpha$. Notably, these results show a vary similar pattern to the perturbation-based attacks: a moderate average ASR with a high variance at small $\alpha$, consistently high ASR for $\alpha = 0.4$ and attack failures for $\alpha = 0.6$ .

Fig. 2 compares the ASR during training between supervised learning and semi-supervised learning with perturbation-based attacks. In supervised learning, the ASR increases gradually from early in training. By contrast, the ASR during semi-supervised learning remains low for many training steps until a point in training at which it rapidly increases to a high ASR.

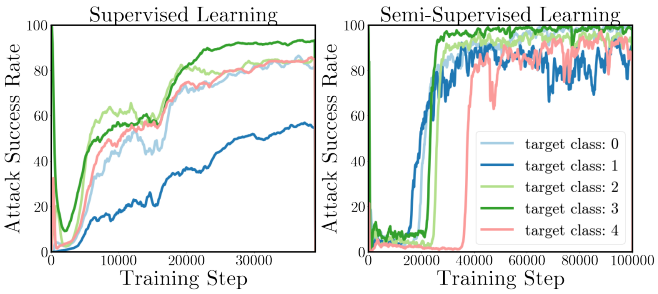



Figure 2: The evolution of ASR during supervised (left) and semi-supervised learning (right).

## 4 Discussion

The two modification-based attacks we used influence two major factors that impact attack performance: the distribution of estimated pseudolabels and the clarity of class-specific features in the poisoned samples. We reason about the observed attack performance by discussing how the strength of the image modifications (perturbation and interpolation) impact these two factors.

When the image modifications are weak or nonexistent, most poisoned samples will receive confident pseudolabels corresponding to the ground truth class label. The poisoned samples will have triggers but they will also have clear target-class-specific features that the network can use for classification, giving the network little reason to rely on the triggers. The success of attacks with weak image modifications vary based on the target class, indicating that some classes have more distinct features that the network can rely on more strongly, weakening the backdoor.

Because the perturbations and interpolations are untargeted, strong modification strength attacks result in high entropy predicted pseudolabels distributed across many classes. Therefore, the network sees triggered samples associated with several classes, leading the network to ignore the trigger as a feature that does not aid in classification. Moderate modification strength attacks are a middle ground in which many poisoned samples will receive confident target class pseudolabels but several samples will be confidently classified as a non-target class or be confusing to the network. These confusing samples will encourage the network to rely more heavily on the triggers, strengthening the backdoor. This analysis suggests that consistently successful backdoor attacks require poison samples that have a pseudolabel distribution heavily concentrated on one class, which can form a weak backdoor, and a subset of poisoned samples that are confusing to the network, which can strengthen the backdoor.

## 5 Conclusion and Practitioner Considerations

We analyzed the effectiveness of backdoor attacks on unlabeled data in semi-supervised learning when the adversary has no control over training labels. This setting requires rethinking the attack development, focusing on the expected distribution of pseudolabels for poisoned samples and the difficulty of recognizing their class-specific features. We gained valuable insight into the impact of backdoor attacks against semi-supervised learning. Our work suggests the possibility of developing a flexible attack that explicitly incorporates a set of samples used to create a weak backdoor and a set of samples used to strengthen the backdoor. To effectively asses the robustness of a semi-supervised learning model or defense to data poisoning attacks, practitioners should consider weak to moderate modification-based attacks with augmentation-robust triggers, and they should investigate attacks across various target classes because, in some settings, target class has an large impact on attack performance (Fig. 1c). Additionally, the ASR during training can remain low for a large portion of training and then spike quickly (rather than increasing gradually as in supervised learning), suggesting that practitioners allow for several training epochs before declaring a model robust to poisoning (Fig 2). Finally, this work highlights the ease with which attacks against semi-supervised learning can be introduced. We observed high average attack success using unperturbed samples that simply have a trigger added. This suggests a serious threat in which adversaries could introduce triggers (either digitally or physically) into the world, these samples could be collected in unlabeled data without ever being reviewed by a human, and a network trained on them could be successfully poisoned.

## REFERENCES


Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. (2020). Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *International Conference on Learning Representations*.

Carlini, N. (2021). Poisoning the unlabeled dataset of {Semi-Supervised} learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1577–1592.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Franci, A., Cordy, M., Gubri, M., Papadakis, M., and Le Traon, Y. (2022). Influence-driven data poisoning in graph-based semi-supervised classifiers. In *2022 IEEE/ACM 1st International Conference on AI Engineering–Software Engineering for AI (CAIN)*, pages 77–87. IEEE.

Gu, T., Dolan-Gavitt, B., and Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Liu, X., Si, S., Zhu, X., Li, Y., and Hsieh, C.-J. (2019). A unified framework for data poisoning attack to graph-based semi-supervised learning. *arXiv preprint arXiv:1910.14147*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Saha, A., Subramanya, A., and Pirsiavash, H. (2020). Hidden trigger backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11957–11965.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., and Goldstein, T. (2021). Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR.

Shejwalkar, V., Lyu, L., and Houmansadr, A. (2022). The perils of learning from unlabeled data: Backdoor attacks on semi-supervised learning. *arXiv preprint arXiv:2211.00453*.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Turner, A., Tsipras, D., and Madry, A. (2019). Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Yan, Z., Wu, J., Li, G., Li, S., and Guizani, M. (2021). Deep neural backdoor in semi-supervised learning: threats and countermeasures. *IEEE Transactions on Information Forensics and Security*, 16:4827–4842.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhao, S., Ma, X., Zheng, X., Bailey, J., Chen, J., and Jiang, Y.-G. (2020). Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14443–14452.

## A Impact of Modification Strength on Pseudolabel Outputs

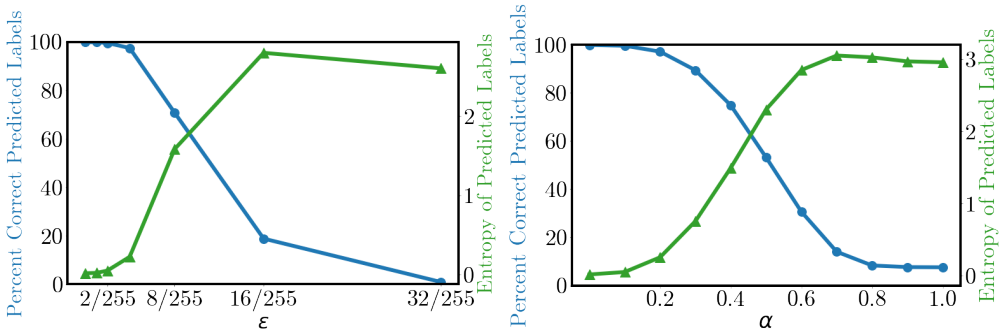
Both of the image modifications used in our experiments, perturbation and interpolation, have the potential to influence the behavior of predicted class outputs. To understand how the strength of the modification impacts the distribution of estimated network outputs, we examine the outputs from a network trained using supervised learning on CIFAR-10 training samples. For perturbation-based attacks, we use Projected Gradient Descent (PGD) adversarial perturbations (Madry et al., 2018), varying the constraint $\epsilon$ on the $\ell_\infty$ norm of the perturbation magnitude. For interpolation-based attacks, we interpolate between our selected target class samples and randomly selected non-target class samples while varying $\alpha$. We apply triggers and weak augmentations to the images to model the poisoned samples in semi-supervised learning. Fig. 3 shows the impact of perturbation and interpolation strength on pseudolabel outputs. The blue line in each plot is the average percentage of modified samples with estimated network outputs that match their ground truth class and the green line is the average entropy of the distribution of class outputs for modified samples. As the perturbation strength increases, fewer poisoned samples are estimated to be the ground truth label and the entropy of the distribution of network outputs increases, indicating the class estimates are distributed more evenly across all class outputs. The same pattern of behavior is seen with the interpolation-based attacks, suggesting similar attack performance between perturbation-based attacks and interpolation-based attacks. While this test is run against a fully trained network, it gives us useful insights for reasoning about the pseudolabels during semi-supervised learning. At low perturbation strengths and small $\alpha$ interpolation values, we expect most poisoned samples have their ground truth classes as pseudolabels. At greater perturbation strength and larger $\alpha$ values, we expect most poisoned samples will not have their ground truth classes as pseudolabels and instead their pseudolabels will be relatively evenly distributed across other classes. These results suggest that both perturbation-based attacks and interpolation-based attacks are successful at impacting pseudolabel behavior.

## B FixMatch Training Details

For the FixMatch implementation, we closely followed the training set up from Sohn et al. (2020). We used a WideResNet-28-2 (Zagoruyko and Komodakis, 2016) architecture, RandAugment (Cubuk et al., 2020) for strong augmentation, and horizontal flipping and cropping for weak augmentation. We used an SGD optimizer with momentum of 0.9, a weight decay of $5 \times 10^{-4}$, and Nesterov momentum. Like Sohn et al. (2020), we used a cosine learning rate decay and quoting from them, we set the "learning rate to $\eta \cos\left(\frac{7\pi k}{16K}\right)$, where $\eta$ is the initial learning rate, $k$ is the current training step, and $K$ is the total number of training steps." We ran 25,000 training epochs and each epoch runs through all the batches of the labeled data. Therefore, with 250 labeled samples, there are four steps per epoch and 100,000 steps total. We report the performance on the exponential moving average of the network parameters. We ensured an even distribution of classes in the labeled data. Additional training parameters are shown in Table 1. We found the following public github repository a good guide to implementing FixMatch: `https://github.com/kekmodel/FixMatch-pytorch`.

## C Adversarial Perturbation Details

For our perturbation-based attacks we used samples that were perturbed using PGD attacks against an adversarially trained network. For $\epsilon = 8, 16, 32/255$ we used perturbed samples provided by the Madry lab whose access locations are specified here: `https://github.com/MadryLab/label-consistent-backdoor-code/blob/main/setup.sh` For $\epsilon = 1, 2, 4/255$ we used perturbed samples generated against a adversarially trained network. The adversarially trained network was a ResNet-50 using $\epsilon = 8/255$ for an $\ell_\infty$ norm. We obtained the weights for the network from the Madry lab: `https://github.com/MadryLab/robustness/#pretrained-models`.




(a) Perturbation-Based Attacks

(b) Interpolation-Based Attacks

Figure 3: Predicted labels of modified samples. (a-b) Percentage of modified training samples with the ground truth class as the estimated label (blue circle line) and the entropy of the distribution of predicted labels (green triangle line) as modification strength is varied. (a) Impact of perturbation-based attacks as $\epsilon$ is varied. (b) Impact of interpolation-based attacks as $\alpha$ is varied.

Table 1: Training parameters for FixMatch

| FixMatch Training Parameters |
|---|
| batch size ($B$): 64 |
| number of epochs: 25000 |
| initial learning rate ($\eta$): 0.03 |
| total number of training steps ($K$): $2^{20}$ |
| poisoning percentage (percentage of entire dataset): $1\%$ (500 samples) |
| number of labeled samples: 250 |
| confidence threshold ($\tau$): 0.95 |
| $\mu$: 7 |
| $\lambda_u$: 1 |

## D Poisoned Sample Details

We used the four corner trigger suggested in Turner et al. (2019) , following the example from `https://github.com/MadryLab/label-consistent-backdoor-code/blob/main/poison_attack.py`, for creating the attack. Fig. 4 shows an example of adversarially-perturbed poisoned images with the four corner trigger. Fig. 5 shows an example of interpolated poisoned images with the four corner trigger.

## E Supervised Learning Details

For supervised learning we also used a WideResNet-28-2 architecture and RandAugment data augmentation during training. We used an SGD optimizer with a momentum of 0.9 and a weight decay of $2 \times 10^{-4}$. We used a multi-step learning rate scheduler that reduced the learning rate by $\gamma = 0.1$ at epochs 40 and 60. To stay consistent with our FixMatch experiments, we report the performance on the exponential moving average of the network parameters.

Figure 4: Poisoned images with increasing perturbation strength $\epsilon$ and the four corner trigger.

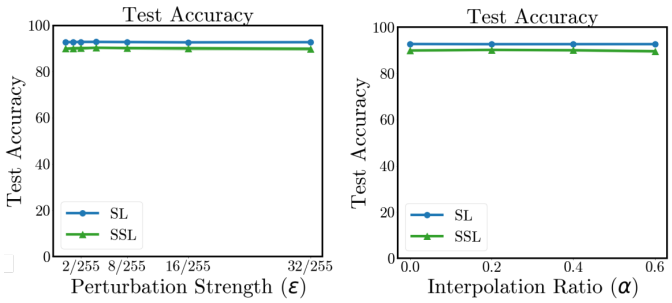
Figure 5: Poisoned images with increasing interpolation ratio $\alpha$ and the four corner trigger.

Table 2: Training parameters for supervised learning

| FixMatch Training Parameters |
|---|
| batch size: 128 |
| number of epochs: 100 |
| initial learning rate ($\eta$): 0.1 |
| poisoning percentage (percentage of entire dataset): $1\%$ (500 samples) |

## F TEST ACCURACY

Fig. 6 shows the test accuracy for the experiments detailed in Fig 1. These results show that the test accuracy remains mostly stable as the modification strength is increased for poisoned samples. There is a slight decrease in test accuracy at the largest interpolation radtio of $\alpha = 0.6$




(a) (b)

Figure 6: (a) Test accuracy for perturbation-based attacks against supervised learning (blue circle line) and semi-supervised learning (green triangle line) with varying $\epsilon$. (b) Test accuracy for interpolation-based attacks.