EFFECTIVE FREQUENCY-BASED BACKDOOR ATTACKS WITH LOW POISONING RATIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attack has been considered a serious threat to deep learning. Although several seminal backdoor attack methods have been proposed, they often required at least a certain poisoning ratio (*e.g.*, 1% or more) to achieve high attack success rate (ASR). However, the attack with a large poisoning ratio may be difficult to evade human inspection or backdoor defenses, *i.e.*, low stealthiness. To tackle the dilemma between high ASR and low stealthiness, we aim to enhance ASR under low poisoning ratio, *i.e.*, pursuing high ASR and high stealthiness simultaneously. To achieve this goal, we propose a novel frequency-based backdoor attack, where the trigger is generated based on important frequencies that contribute positively to the model prediction with respect to the target class. Extensive experiments on four benchmark datasets (CIFAR-10, CIFAR-100, GTSRB, Tiny ImageNet) verify the effectiveness and stealthiness of the proposed method under extremely low poisoning ratios. Specifically, with only 0.01% poisoning ratio, our attack could achieve the ASR of 80.51%, 51.3%, 76.3%, and 87.2% on above four datasets, respectively, while the ASR values of most state-of-the-art (SOTA) attack methods are close to 0. Meanwhile, our method could well evade several SOTA backdoor defense methods, *i.e.*, the ASR values are not significantly affected under defense.

1 INTRODUCTION

Training high-performance deep learning models often require large-scale training data and sufficient computing resources. To address these two obstacles, one feasible solution is to buy or freely download a third-party large-scale dataset or a pre-trained model. However, the unverified dataset or pre-trained model may bring serious security threats to deep learning models. One typical threat is backdoor attack, which aims to mislead the model training through manipulating some training samples, such that the trained model performs normally on benign samples, while predicts any poisoned sample to the target class.

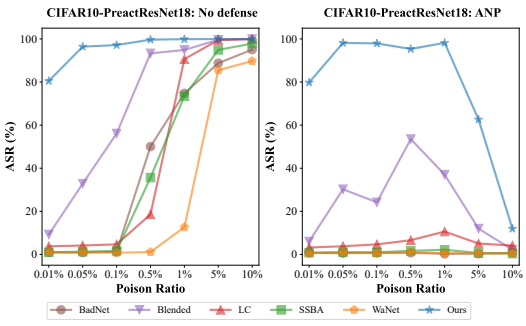


Figure 1: Examples of backdoor attack with low poisoning ratios. **(Left)**: no defense; **(Right)**: under the ANP defense Wu & Wang (2021a).

There are two requirements of a successful backdoor attack, including: *effectiveness* (measured by attack success rate (ASR)), and *stealthiness* (measured by the ability to bypass human inspection and backdoor defense). Several seminal backdoor attacks have shown the superior performance of high effectiveness and high stealthiness simultaneously. However, we observed that these attacks were often evaluated with 1% or higher poisoning ratio (*i.e.*, the percentage of poisoned samples in the training set). When a lower poisoning ratio is adopted, we find that several state-of-the-art attacks give much lower ASR scores, as shown in Fig. 1. On the other hand, the attack with high poisoning ratio may sacrifice the stealthiness. First, it may arouse human suspicion, as lots of training samples are incorrectly labeled to the same class. Second, as observed and analyzed in BackdoorBench Wu et al. (2022), the attack with higher poisoning ratio is more likely to be

detected or erased by backdoor defenses, as the difference between backdoored and benign models may be large enough to be identified by the defender.

To address the above dilemma between the effectiveness and stealthiness of existing backdoor attacks, this work aims to improve the ASR with very low poisoning ratios, such that high ASR and high stealthiness could be ensured simultaneously. To achieve this goal, we propose a novel frequency-based backdoor attack method according to the frequency importance. First, we present a novel technique to measure the importance of each frequency spectrum in the frequency domain, such that the frequency spectrum with positive contribution to the model prediction of one image could be identified. Second, we pick out the top-k important frequency spectrum of one benign image of the target class (also called as *target anchor image*) as the frequency trigger. Then, given one benign image from other classes, we insert the frequency trigger into its frequency spectrum map, *i.e.*, replacing its original frequency spectrum by the trigger at the corresponding k frequency indices. This obtained frequency spectrum map is then transformed to the RGB image via the inverse discrete Fourier transform (IDFT). The label of this new RGB image is changed to the target class, such that we generate one poisoned sample. Utilizing the poisoned samples generated above, we can set up a data poisoning based backdoor attack against deep learning models. Extensive experiments on several benchmark datasets demonstrate the superior performance on both effectiveness and stealthiness of the proposed attack method, especially in the case of extremely low poisoning ratios. For example, with only 0.01% poisoning ratio, the ASR of our method achieves 80.51%, while those of other compared methods are close to 0 (see Fig. 1-Left), which demonstrates the better effectiveness of our method. Moreover, even under the ANP defense, our method still keeps high ASR with 1% or smaller poisoning ratios, while the ASR values of other methods are significantly downgraded when the poisoning ratio passes 0.5% (see Fig. 1-Right). Besides, we can adjust the k value in the frequency trigger to control the visual distortion of each individual poisoned image, and the extremely low poisoning ratio means high possibility to evade human inspection. Consequently, these aspects guarantee the high stealthiness of the proposed method.

The main contributions of this work are three-fold. **1)** We present a novel technique to measure the contribution of each frequency component to the model prediction on one image. **2)** We propose an effective and stealthy frequency-based backdoor attack method, which could achieve high ASR with extremely low poisoning ratio. **3)** Extensive experiments on four benchmark datasets demonstrate the superior performance of the proposed method to several state-of-the-art backdoor attack methods, even under state-of-the-art backdoor defenses.

2 RELATED WORK

Backdoor attacks As a serious threat to deep learning, Backdoor attacks can be roughly classified into two categories, including data poisoning attack(Gu et al., 2019) and training controllable attack (Shumailov et al., 2021). For data poisoning attacks, adversaries can inject triggers into some samples and modify their labels (Gu et al., 2019; Chen et al., 2017). In this line of research, a series of methods aiming to improve the attack performance and evade human inspection have been proposed, including sample agnostic attackBarni et al. (2019), Li et al. (2021b), sample-specific attackSouri et al. (2021), invisible backdoor attack Li et al. (2020a) and clean label backdoor attack (Shafahi et al., 2018; Zhao et al., 2020). For training controllable attacks, attackers can control both the training process and the dataset. Typical training controllable backdoor attacks include Input-awareNguyen & Tran (2020), WaNet (Nguyen & Tran, 2021).

Backdoor defenses In general, defense methods have three categories **pre-training**, **in-training** and **post-training**. **(1)pre-training** methods aim to remove or distort poisoned samples before training. Borgnia et al. (2020) suggested that methods such as data augmentation could be used to attenuate or eliminate the effects of backdoors during training. **(2)in-training** methods are designed to inhibit the learning of backdoor in the training process. Li et al. (2021a) proposed Anti-backdoor learning (ABL) to make use of the difference in training loss decline rate between backdoor samples and clean samples in the training process and to distinguish and unlearn backdoor samples. Based on the intuition that the feature differences of self-supervised learning can be used to make it difficult for the model to learn backdoor samples and semi-supervised learning data, Huang et al. (2022) propose a backdoor defense method by decoupling the training process. **(3)post-training** aims to remove or eliminate the backdoor in a backdoored model based on some special properties of backdoors. The

adversarial neuron pruning (ANP) defense Wu & Wang (2021b) and the channel Lipschitzness based pruning (CLP) method Zheng et al. (2022) snipped out the possible backdoor neurons they found to mitigate the impression of backdoor attack. The activation clustering (AC) method Chen et al. (2019) and the spectral signatures (Spectral) method Tran et al. (2018) use the characteristics of clean samples and backdoor samples to screen the backdoor samples. Also, the neural cleanse (NC) method Wang et al. (2019) and Tabor method Guo et al. (2020) learn the reversed trigger from the backdoor model and relearn the sample with Reversed Trigger. The influence of backdoor neurons was eliminated by guiding the model with clean samples in the fine-pruning (FP) defense Liu et al. (2018) and the neural attention distillation (NAD) Li et al. (2020b).

Frequency-based Backdoor attacks To utilize the advantages of frequency domain that small pixel-wise perturbations can disperse across the entire image and CNNs are able to learn from images' frequency-domain features, Wang et al. (2021) proposed FTrojan which transforms the images from RGB to HUV and injects small perturbation in at mid- and/or high-frequency components of UV channel. Through Fourier Heatmaps Yin et al. (2019), Hammoud & Ghanem (2021) proposed a frequency backdoor attack that injects triggers at the top-k sensitive frequencies. Rather than using perturbation as trigger, Kwon & Kim (2022) designed blind-watermark backdoor method in which the frequency components of poisoning samples are generated by synthesizing the components of clean samples with a specific image. Besides, Feng et al. (2022) developed the Frequency-Injection based Backdoor Attack (FIBA) method that triggers are injected in amplitude spectrum which maintains the phase spectrum information. By analyzing the frequency spectrum of poisoning samples, Zeng et al. (2021) discovered that most triggers used in backdoor attacks exhibit high-frequency artifacts, so they proposed an evaluation metric called Backdoor data Detection Rate (BDR) to separate clean samples and the poisonings. Then, they proposed a low-frequency backdoor attack that utilized adversarial attack to find a UAP trigger and then put it into a low-pass filter to acquire a smooth trigger. However, these frequency-based attacks haven't analyzed the importance of frequency components.

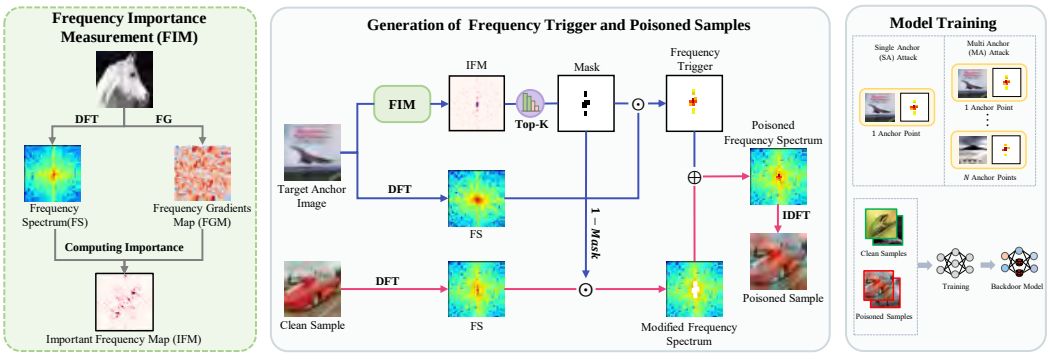


Figure 2: The pipeline of Important Frequency Backdoor Attack.

3 Method

3.1 Notation and threat model

Notation We denote $\mathcal{D}_o = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^N$ as the benign training dataset with S classes and N samples where $\boldsymbol{x}_j \in \mathbb{R}^{H\times W\times C}$ is the jth clean image of size $H \times W \times C$ and $y_j \in \{1, \ldots, S\}$ is the jth label. In the rest of this paper, we assume $C = 3$. Let $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^S$ be the classifier whose input is the original image $\boldsymbol{x}$ and output is a vector of logits. Let $\mathbb{C}$ be the space of complex numbers. We denote $\tilde{f} : \mathbb{C}^{H\times W\times C} \to \mathbb{R}^S$ as the corresponding frequency domain classifier, where the input is the frequency spectrum map of the input image $\boldsymbol{x}$, *i.e.*, $\tilde{\boldsymbol{x}} = \mathfrak{F}(\boldsymbol{x})$ with $\mathfrak{F}$ being the channel-wise Discrete Fourier Transform (DFT) operator Brigham & Morrow (1967). And, we have

$$\tilde{f}(\tilde{\boldsymbol{x}}) = f(\boldsymbol{x}) = f\big(\mathfrak{F}^{-1}(\tilde{\boldsymbol{x}})\big), \tag{1}$$

where $\mathfrak{F}^{-1}$ is the channel-wise Inverse Discrete Fourier Transform (IDFT) operator, which means $\mathfrak{F}^{-1}(\tilde{x}) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\tilde{x}(h,w)e^{-2\pi i(\frac{uh}{H}+\frac{vw}{W})}$

Threat Model We consider the data poisoning based backdoor attack, where the attacker only has the access to the training dataset D_o, which can be divided into two parts $\mathcal{D}_c$ and $\mathcal{D}_p$. The attacker manipulates $\mathcal{D}_p$ to $\mathcal{D}_p^* = \{(\boldsymbol{x}_j^*, t) \mid \boldsymbol{x}_j^* = \mathcal{P}(\boldsymbol{x}_j, \boldsymbol{\eta}), (\boldsymbol{x}_j, y_j) \in \mathcal{D}_p\}$, where the operator $\mathcal{P}(\boldsymbol{x}, \boldsymbol{\eta})$ injects a trigger $\boldsymbol{\eta}$ into the image $\boldsymbol{x}$, and t denotes the target label. Then, if the model f is trained on the poisoned training dataset $\mathcal{D}_{poison} = \mathcal{D}_c \cup \mathcal{D}_p^*$, f is expected to predict the poisoned image $\boldsymbol{x}^*$ as the target label, while predict the benign image $\boldsymbol{x}$ correctly.

3.2 Frequency importance measure

Frequency Gradient Given an image $\boldsymbol{x}$ and the corresponding frequency spectrum map $\tilde{\boldsymbol{x}}$, we denote its frequency gradient map *w.r.t.* class s as $\mathcal{G}_s = \frac{\nabla \tilde{f}_s}{\nabla \tilde{\boldsymbol{x}}} \in \mathbb{C}^{H\times W\times C}$. According to Wu et al. (2022), $\mathcal{G}_s$ is calculated as follows:

$$\begin{aligned}\mathcal{G}_s(u,v,c) = \frac{\partial \tilde{f}_s(\tilde{\boldsymbol{x}})}{\partial \tilde{\boldsymbol{x}}(u,v,c)} &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sum_{c'=0}^{C}\frac{\partial f_s(\boldsymbol{x})}{\partial \boldsymbol{x}(h,w,c')}\cdot\frac{\partial \boldsymbol{x}(h,w,c')}{\partial \tilde{\boldsymbol{x}}(u,v,c)} \\ &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\frac{\partial f_s(\boldsymbol{x})}{\partial \boldsymbol{x}(h,w,c)}e^{2\pi i(\frac{uh}{H}+\frac{vw}{W})},\end{aligned} \tag{2}$$

where the third equation holds since we utilize channel-wise DFT in (1).

Frequency Importance To measure the contribution of each entry in $\tilde{x}$ to the prediction $\tilde{f}_k(\tilde{x})$, we propose to approximately factorize the objective function $\tilde{f}_s(\tilde{\boldsymbol{x}})$ by Taylor's expansion. Specifically, $\tilde{f}_s(\tilde{\boldsymbol{x}})$ can be approximated by

$$\tilde{f}_s(\tilde{\boldsymbol{x}}) \simeq \tilde{f}_s(\bar{\boldsymbol{x}}) + \sum_u\sum_v\sum_c \mathcal{G}(u,v,c)\left[\tilde{\boldsymbol{x}}(u,v,c) - \bar{\boldsymbol{x}}(u,v,c)\right], \tag{3}$$

where $\bar{\boldsymbol{x}}$ is a frequency spectrum map. In this paper, we set $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{j=1}^{N}\tilde{\boldsymbol{x}}_j$.

Since $\tilde{f}_s(\tilde{\boldsymbol{x}})$ is real valued, we can further simplify 3 as

$$\tilde{f}_s(\tilde{\boldsymbol{x}}) \simeq \mathfrak{Re}\left(\tilde{f}_s(\bar{\boldsymbol{x}})\right) + \sum_u\sum_v\sum_c \mathfrak{Re}\left(\mathcal{G}(u,v,c)\left[\tilde{\boldsymbol{x}}(u,v,c) - \bar{\boldsymbol{x}}(u,v,c)\right]\right), \tag{4}$$

where the function $\mathfrak{Re}$ keep the real part of the complex number, *i.e.*, $\mathfrak{Re}(a+bi) = a$.

Equation 4 suggests that the contribution of $\tilde{x}(u,v,c)$ to $\tilde{f}_s(\tilde{x})$ can be approximated as $\mathfrak{Re}\left(\mathcal{G}(u,v,c)\left[\tilde{\boldsymbol{x}}(u,v,c) - \bar{\boldsymbol{x}}(u,v,c)\right]\right)$. Therefore, we define the importance frequency map (IFM) $\phi \in \mathbb{R}^{H\times W}$ as

$$\phi(u,v) = \mathfrak{Re}\left(\sum_c \mathcal{G}(u,v,c)\left[\tilde{\boldsymbol{x}}(u,v,c) - \bar{\boldsymbol{x}}(u,v,c)\right]\right). \tag{5}$$

Remark that a positive $\phi(u,v)$ indicates positive impact of $\sum_c \tilde{x}(u,v,c)$ on the prediction results.

3.3 Poisoned sample generation

Frequency Trigger Here we generate a frequency trigger related to the target class t according to the frequency importance map. Specifically, we first choose an benign image $\boldsymbol{x}_a$ from the target class t, which is also called target anchor image, then we calculate its frequency spectrum map $\tilde{\boldsymbol{x}}_a$ and the frequency importance map *w.r.t.* class t, *i.e.*, $\mathcal{FI}_t$. We denote $\mathbb{I}_k$ as the index set of the top-k largest entries in ϕ. Then, we define a frequency importance mask $\mathcal{M} \in \{0,1\}^{H\times W}$ as follows

$$\mathcal{M}(u,v) = \begin{cases}1, & (u,v) \in \mathbb{I}_k \\ 0, & \text{otherwise.}\end{cases} \tag{6}$$

Further, we define a frequency trigger $\mathcal{T}_a$ corresponding to $\boldsymbol{x}_a$, as follows

$$\mathcal{T}_a = \mathcal{M} \odot \tilde{\boldsymbol{x}}_a, \tag{7}$$

where $\odot$ is the entry-wise multiplication operator for each channel. $\mathcal{T}_a$ contains the top-k importance frequency spectrum of $\boldsymbol{x}_a$ *w.r.t.* the target class t.

Poison sample For each benign image $\boldsymbol{x}$ in $\mathcal{D}_p$, we first obtain its frequency spectrum $\tilde{\boldsymbol{x}}$ by DFT. Then, we insert the frequency trigger $\mathcal{T}_a$ into $\tilde{\boldsymbol{x}}$ to obtain a poisoned frequency spectrum $\tilde{\boldsymbol{x}}^*$,

$$\tilde{\boldsymbol{x}}^* = \mathcal{T}_a + (1 - \mathcal{M}) \odot \tilde{\boldsymbol{x}}. \tag{8}$$

Note that here we simply replace the original frequency spectrum at the indices $\mathbb{I}_k$ of $\tilde{\boldsymbol{x}}$ by the corresponding spectrum in $\mathcal{T}_a$, and other combination modes could be explored in future. Finally, we obtain a poisoned image $\boldsymbol{x}^* = \mathfrak{F}^{-1}(\tilde{\boldsymbol{x}}^*)$, and modify its label as the target class t.

3.4 Important frequency based backdoor attack

Once a poisoned training dataset $\mathcal{D}_{poison}$ is generated as described above, it can be released to the developer to train a backdoored model, which is expected to be fooled the poisoned sample in the testing stage. This backdoor attack method is called Important Frequency based Backdoor Attack (IFBA). Furthermore, according to the number of adopted target anchor images when generating the poisoned sample, we denote two variants of IFBA as IFBA-SA (*i.e.*, using a single anchor image) and IFBA-MA (*i.e.*, using multiple anchor images). Note that there will be multiple frequency triggers corresponding to different anchor images, and one of them will be randomly chosen to generate one poisoned sample, similar to dynamic triggers. The performance of these two variants will be evaluated later.

4 Experiments

4.1 Experimental setting

Datasets and Models. We evaluate all baselines and the proposed method on four popular datasets including CIFAR-10, CIFAR100, Krizhevsky et al. (2009), GTSRB Houben et al. (2013) and Tiny ImageNet Le & Yang (2015). For each dataset and method, we conduct experiments using two classical models, i.e. PreAct-ResNet18 He et al. (2016) and VGG19 Simonyan & Zisserman (2014).

Baselines and Setups. To benchmark our method with the pioneering attack/defense methods, we follow the work Wu et al. (2022) to compare our method with 8 attack methods (BadNets Gu et al. (2019), Blended Chen et al. (2017), LC Shafahi et al. (2018), SIG Barni et al. (2019), LF Zeng et al. (2021), SSBA Li et al. (2021b), Input-aware Nguyen & Tran (2020), WaNet Nguyen & Tran (2021)) and 9 defense methods (Fine-tuning (FT), FP Liu et al. (2018), NAD Li et al. (2020b), NC Wang et al. (2019), ANP Wu & Wang (2021a), AC Chen et al. (2019), Spectral Tran et al. (2018), ABL Li et al. (2021a), DBD Huang et al. (2022)). For frequency-based methods, only the LF method was compared in this paper since other mentioned frequency-based methods are still not open-sourced yet. We consider 5 poisoning ratios, *i.e.*, 0.01%, 0.05%, 0.1%, 0.5%, and 1%, for each attack-defense pair on all datasets and models. In all experiments, only entries in ϕ with positive value are considered for trigger generation. For the top-k operation, we set the value of k as 3, 10, 100, and full, *i.e.*, the number of all positive entries in ϕ. We refer readers to **Appendix** A.2 for more details.

Evaluation metrics. To measure the performance of all methods, we report the attack success rate (ASR) (*i.e.*, the probability of classifying a poisoned test data to the target label) and the clean accuracy (CA) (*i.e.*, the probability of classifying a benign test data to the correct label) in all experiments.

4.2 Attack evaluation without defense

Comparisons with SOTA methods Here we set the top-k value as the number of all positive contributed frequency in the importance frequency map ϕ for our methods. The evaluations of 8 compared methods and our methods on CIFAR-10 dataset and PreactResNet18 model are summarized

Table 1: Evaluations on CIFAR-10 and PreactResNet18. IFBA-MA(3) indicates 3 target anchor images in our method.

Attack ↑	0.01%		0.05%		0.1%		0.5%		1%	
	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
BadNet	93.74	0.92	93.89	0.84	93.61	1.23	93.76	50.06	93.14	74.73
Blended	93.97	9.26	93.69	32.84	93.80	56.11	93.68	93.30	93.76	94.88
Input-aware	91.52	7.46	91.25	70.67	91.94	47.32	90.66	56.06	91.74	79.18
LC	**94.20**	3.79	93.73	4.13	93.82	4.67	93.66	18.61	93.78	90.63
LF	93.87	2.29	93.92	5.20	93.55	12.72	**93.85**	75.09	93.56	86.46
SIG	93.85	2.84	93.68	35.18	93.73	41.27	93.80	82.43	93.82	83.40
SSBA	93.76	1.00	93.65	1.32	93.89	1.62	93.41	35.67	93.43	73.44
WaNet	90.96	1.06	91.28	1.08	92.18	0.78	91.27	1.12	90.65	12.63
IFBA-SA	94.01	**80.51**	93.79	**96.41**	93.77	**97.17**	93.81	**99.66**	93.69	**99.89**
IFBA-MA(3)	93.68	36.59	**93.95**	80.91	**93.95**	92.38	93.61	98.61	**93.97**	99.58
IFBA-MA(5)	93.85	44.86	93.93	78.61	93.83	93.18	93.72	99.24	93.56	99.58

in Table 1. Our method IFBA-SA shows the highest ASR and competitive CA at all poisoning ratios, while IFBA-MA takes the second rank. Especially, even with 0.01% poisoning ratio, the ASR of IFBA-SA is up to 80.5%, while those of all compared methods are lower than 10%. These evaluations fully demonstrate the superior effectiveness of our attack method to SOTA attacks. Note that due to the space limit, evaluations on other datasets are shown in **Appendix**.

Effect with different top-k values We further evaluate the effect of different top-k values in our trigger, with $k = 3, 10, 100,$ full, respectively. As shown in Figure 3, attack higher top-k values always gives higher ASR values in all cases. It is reasonable that higher top-k values means that the trigger will contain more important information of the target anchor image(s), such that it is easier to learn the stable mapping from the trigger to the target class.

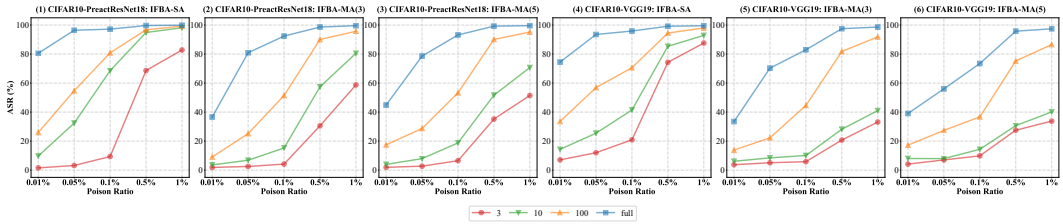


Figure 3: Results measured by ASR of different top-k values in our methods (IFBA-SA, IFBA-MA(3), IFBA-MA(5)), on PreactResNet and VGG19 model structures.

4.3 Attack Evaluation under Defenses

Comparisons with SOTA methods Here we set the top-k value as the number of all positive contributed frequency in the importance frequency map ϕ for our methods. The evaluations of our attacks and 8 compared attacks against 9 SOTA backdoor defense methods are shown in Figure 4. Our attacks give much higher ASR values than all compared attacks, under all defenses. Especially under the ANP defense, only our IFBA-SA attack achieves high ASR at all poisoning ratios. Besides, it is notable that the ASR values of our methods are not significantly affected by most defenses, compared with those of no defense (see the top-left sub-figure), which demonstrates the superior ability to evade defenses. We note that, under ABL and DBD, there are obvious degradation of our attacks. However, as shown in Figure 5, the corresponding CA values are also significantly decreased, illustrating that ABL and DBD don't provide successful defenses against our attacks. Above evaluations verify the superior ability to evade backdoor defenses of our method to other compared attacks. Note that due to the space limit, evaluations on other datasets are shown in **Appendix**.

IFBA-SA *vs.* IFBA-MA with different top-k values We further compare the attack performance under defenses between IFBA-SA and IFBA-MA, with different top-k values, as shown in Figure 5.

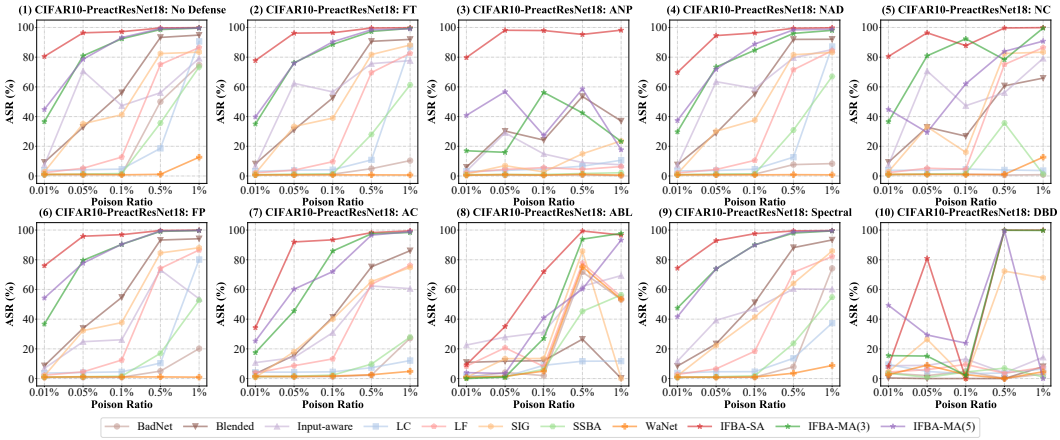


Figure 4: Attack performance under 9 defenses measured by ASR between our attack methods and 8 compared attack methods.

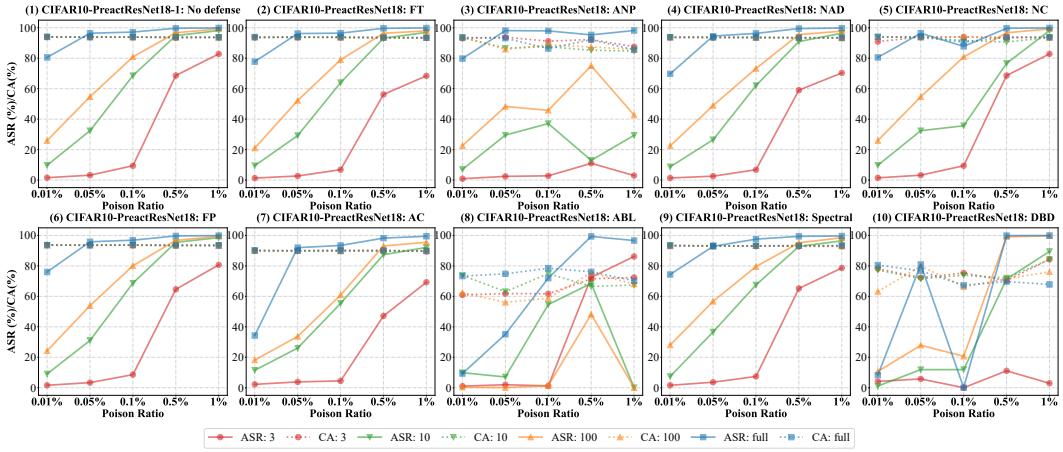
As shown in the top 2 rows, IFBA-SA still performs very well under most defenses, with high ASR and high CA values, especially with top-k being full. The only 2 exceptions are ABL and DBD, both ASR and CA of IFBA-SA are decreased, demonstrating not satisfied defense. The evaluations of IFBA-MA are shown in the bottom 2 rows. The trends are similar with IFBA-SA, with the only difference that ANP shows good defense against IFBA-MA.

4.4 Analysis and Visualization

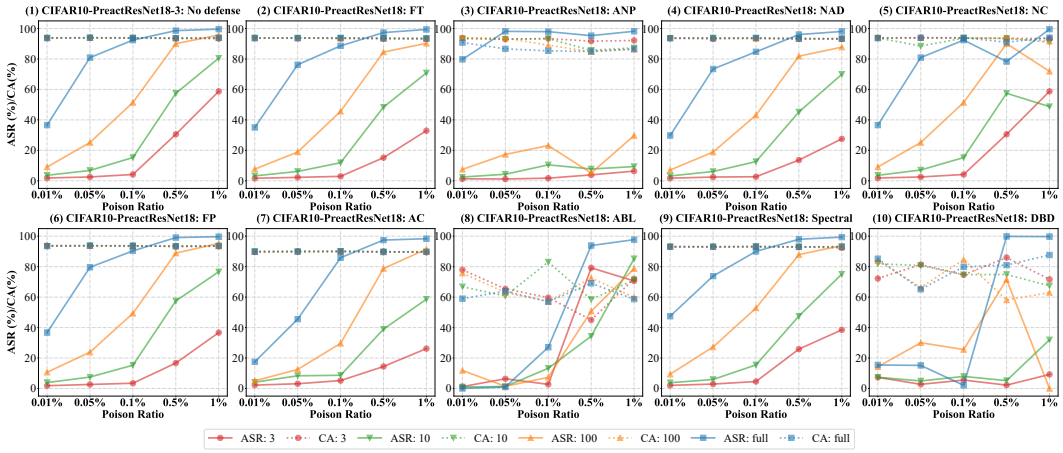
Visual effects of different top-k values in the frequency trigger The ASR results and some poisoned images of different top-k values with different poisoning ratios are shown in Figure 6. In all the four datasets, it's clear that with the increasing of poisoning ratio, ASR of all the four top-k values increase. When the poisoning ratio is 1%, even the lowest ASR of top-3 was higher than 80% in the dataset of CIFAR-10. And higher top-k value led to higher ASR under the same poisoning ratio. When top-k is set as full, even the lowest ASR at 0.01% poisoning ratio achieves approximately 60% ASR on CIFAR-100. In short, increasing both poisoning ratio and top-k value are beneficial to achieve higher ASR. However, as shown in the bottom rows, higher top-k values will cause more obvious visual distortion compared to the benign image. Thus, there is a trade-off between the visual stealthiness and ASR *w.r.t.* the top-k value. However, the extremely low poisoning ratio guarantees the stealthiness of our attack under human inspection, even with high visual distortion on individual poisoned samples.

T-SNE visualization To visualize the performance of backdoor models, we utilize T-SNE method to observe the distribution of clean samples and poisoned samples, whose quantities are 5000 and 500, respectively. Figure 7 presents the T-SNE results of IFBA-SA_full attack under five different poisoning ratios, in which the red pentacle represents the anchor point. It can be seen that even at the lowest poisoning ratio of 0.01%, poisoned samples located closely to the target samples. With the increase of poisoning ratio, they migrate to each other more closely. At the same time, distributions of clean samples were almost not influenced.

Grad-CAM Detection In this part, we provide the Grad-CAM results about SA attack on CIFAR-10 with PreactResNet18. As shown in Figure 8, the first column is original images. The second column is Grad-CAM of a clean sample passing through the clean model. And the rest of them are Grad-CAM the poisoning samples passing through corresponding backdoor models. From left to right, their top-k are '3', '10', '100', and 'full', respectively. It can be seen that our IFBA-SA attack do not make the attention area of models deviate from the object, when compared to the clean model.



(a) The defense results of IFBA-SA.



(b) The defense results of IFBA-MA(3).

Figure 5: Attack performance under defenses measured by ASR and CA of our attack methods, with different top-k values: **(a)** IFBA-SA; **(b)** IFBA-MA(3).

5 Conclusion

This work has proposed an effective and stealthy frequency based backdoor attack method. Based on a novel technique to measure the frequency importance to the model prediction, we designed a frequency-based trigger, which contained the important frequency spectrum of the target anchor images. By inserting this trigger into the frequency spectrum map of benign images from other classes, we generate stealthy poisoned images with slight visual distortion. Extensive evaluations on several benchmark datasets demonstrated the superior performance on both attack effectiveness and stealthiness of the proposed attack method to 8 SOTA backdoor attack methods, even under 9 SOTA backdoor defense methods, especially with extremely low poisoning ratios (*e.g.*, 0.01%).

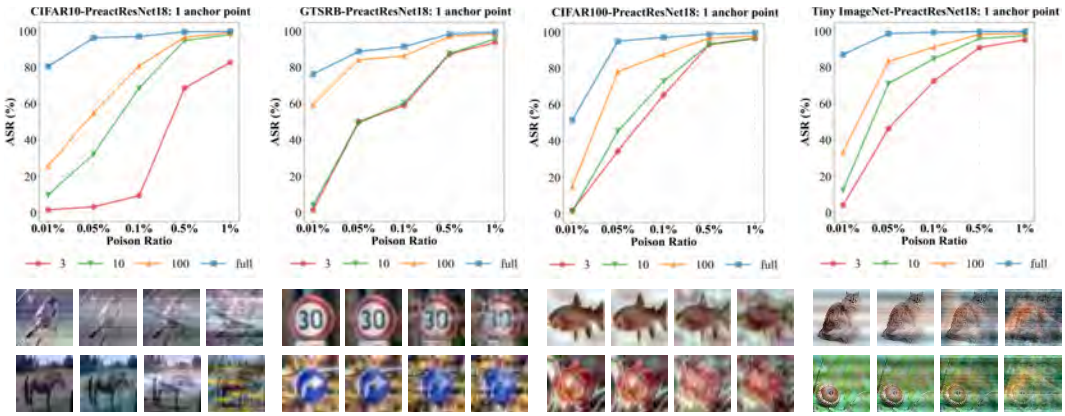


Figure 6: ASR variation of different top-k value to different poisoning ratio in different datasets and corresponding poisoned image of different top-k value

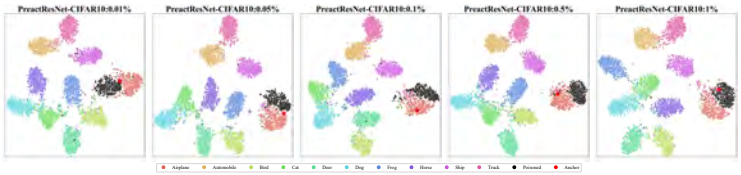


Figure 7: T-SNE results of five different poisoning ratios for PreactResNet18 on CIFAR-10

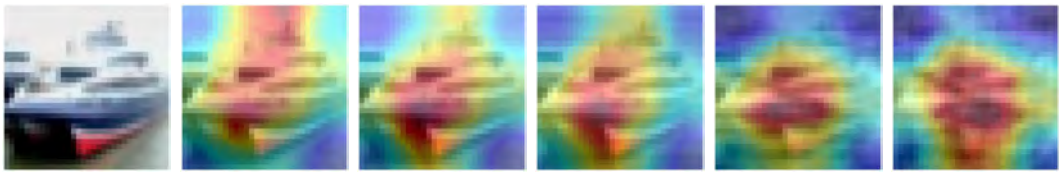


Figure 8: Grad-CAM results of benign image and four poisoned images.

REFERENCES

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105. IEEE, 2019.

Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping, Micah Goldblum, Tom Goldstein, and Arjun Gupta. Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. *arXiv preprint arXiv:2011.09527*, 2020.

E. O. Brigham and R. E. Morrow. The fast fourier transform. *IEEE Spectrum*, 4(12):63–70, 1967. doi: 10.1109/MSPEC.1967.5217220.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *The AAAI Conference on Artificial Intelligence Workshop*, 2019.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20876–20885, 2022.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2019.

Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Towards inspecting and eliminating trojan backdoors in deep neural networks. In *IEEE International Conference on Data Mining*, 2020.

Hasan Abed Al Kader Hammoud and Bernard Ghanem. Check your other door! establishing backdoor attacks in the frequency domain. *arXiv preprint arXiv:2109.05507*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, number 1288, 2013.

Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Hyun Kwon and Yongchul Kim. Blindnet backdoor: Attack on deep neural network using blind watermark. *Multimedia Tools and Applications*, 81(5):6217–6234, 2022.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

Shaofeng Li, Minhui Xue, Benjamin Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 2020a.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2020b.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34, 2021a.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16463–16472, 2021b.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pp. 273–294. Springer, 2018.

Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33, 2020.

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eEn8KTtJOx.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A Erdogdu, and Ross J Anderson. Manipulating sgd with data ordering attacks. *Advances in Neural Information Processing Systems*, 34:18021–18032, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Hossein Souri, Micah Goldblum, Liam Fowl, Rama Chellappa, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *arXiv preprint arXiv:2106.08970*, 2021.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems Workshop*, 2018.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723. IEEE, 2019.

Tong Wang, Yuan Yao, Feng Xu, Shengwei An, and Ting Wang. Backdoor attack through frequency domain. *arXiv preprint arXiv:2111.10991*, 2021.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Chao Shen, and Hongyuan Zha. Backdoorbench: A comprehensive benchmark of backdoor learning. In *NeurIPS 2022 Track Datasets and Benchmarks*, 2022.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34, 2021a.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34, 2021b.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16473–16481, 2021.

Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14443–14452, 2020.

Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. *arXiv preprint arXiv:2208.03111*, 2022.

A Additional information of experiments

A.1 FSM

FSM visualization Frequency Saliency Map (FSM) is a visualization tool in the frequency domain that can indicate which frequencies affect classification result (Wu et al., 2022). We scan our backdoor models, including IFBA-SA, IFBA-MA(3) and IFBA-MA(5) by FSM, to check their stealthiness. The results were shown in Figure 9 where the warm level of color represent the contribution of frequencies at different locations. The first row was the result of IFBA-SA with increasing top-k value, the second and third rows were results of IFBA-MA(3) and IFBA-MA(5) with different top-k value, respectively. Among them, all the poisoning ratios were set as 0.1%. In IFBA-SA, locations of the frequency trigger can be detected by FSM, as shown as the red point in the first row of Figure 9. While the distribution of warm color in the results of IFBA-MA(3) and IFBA-MA(5) were both disperse, which is hard to recognize the locations of frequency triggers. The MA methods show good stealthiness, comparing to that of the SA method.

A.2 Experiment Details

Settings for baselines In all experiments, we adopt the default settings for attack baselines and defense baselines provided in Wu et al. (2022). Here, we provide a summary of general settings and refer readers to Wu et al. (2022) for the details of each baseline method. For all baselines, we use SGD optimizer with momentum 0.9, weight decay 0.0005. For CIFAR10 and CIFAR100, we

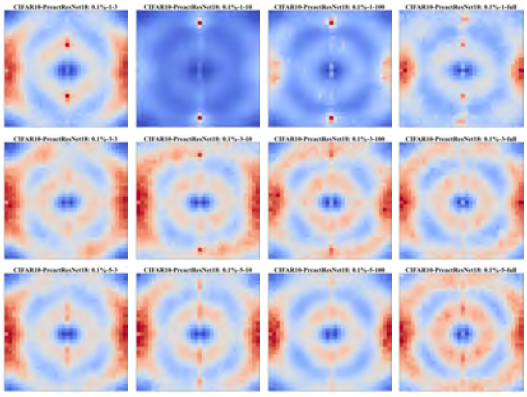


Figure 9: The FSM visualizations for IFBA-SA, IFBA-MA(3), IFBA-MS(5) with different top-k on CIFAR-10.

train models for 100 epochs with CosineAnnealingLR schedule. for GTSRB, we train models for 50 epochs and CosineAnnealingLR schedule. For TinyImageNet, we train models for 200 epochs and ReduceLROnPlateau schedule. We adopt batch size 128 for attack and batch size 256 for defense

Settings for our method In anchor points selection, we choose 10 samples with high classification confidence from the target class, such as 'Airplane'. For SA attack, we always choose the first images of 10 as the anchor point. As for MA, we randomly select one from top-k anchor points at each time.

A.3 Results of IFBA on CIFAR-100

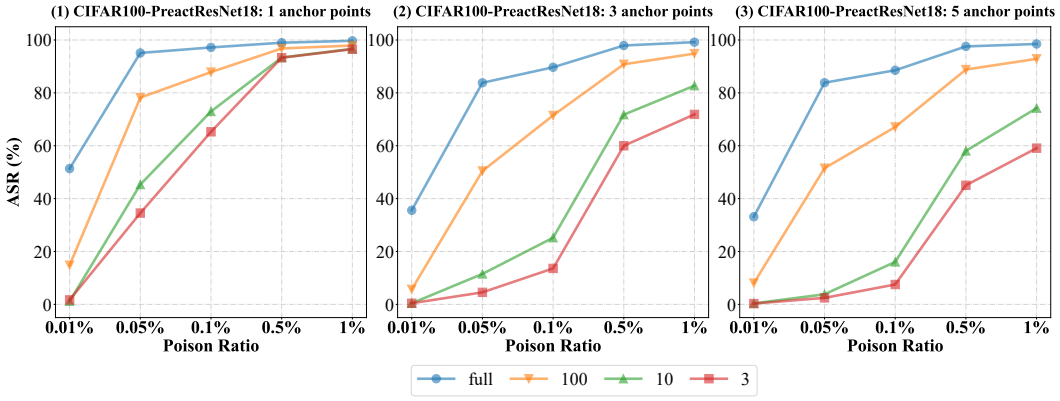


Figure 10: The attack and defense results of SA and MA on CIFAR100 with PreactResNet18

A.4 Results of IFBA on GTSRB

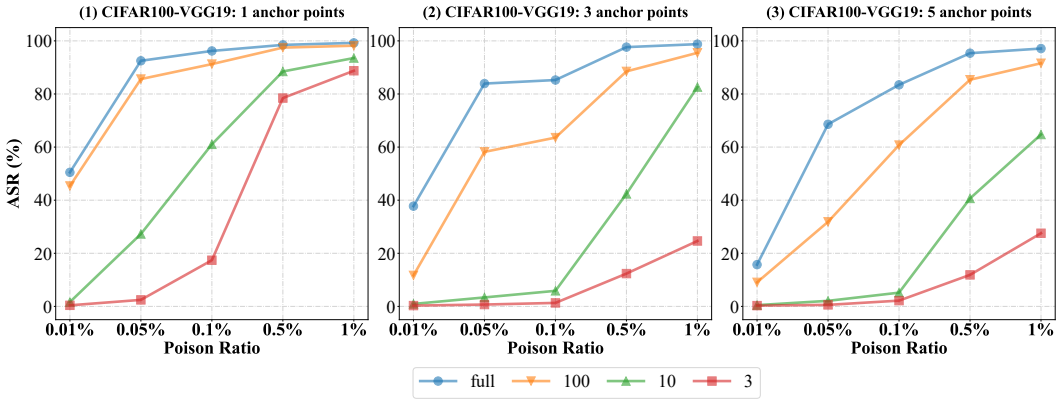


Figure 11: The attack results of SA and MA on CIFAR100 with PreactResNet18

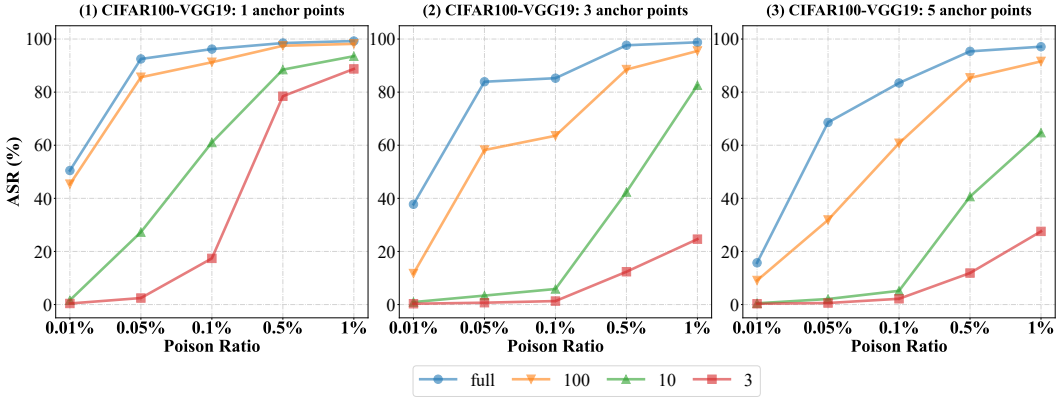


Figure 12: The attack results of SA and MA on CIFAR100 with VGG19

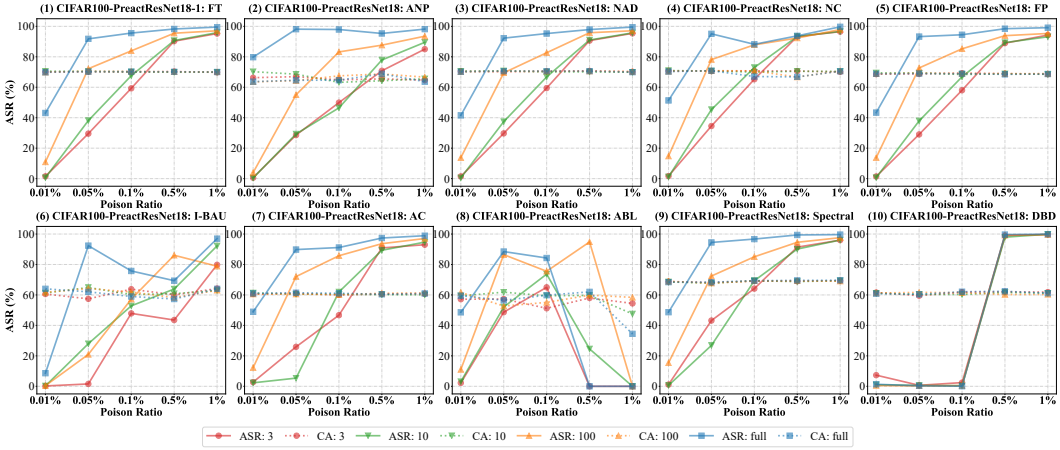


Figure 13: The defense results of SA and MA on CIFAR100 with PreactResNet18

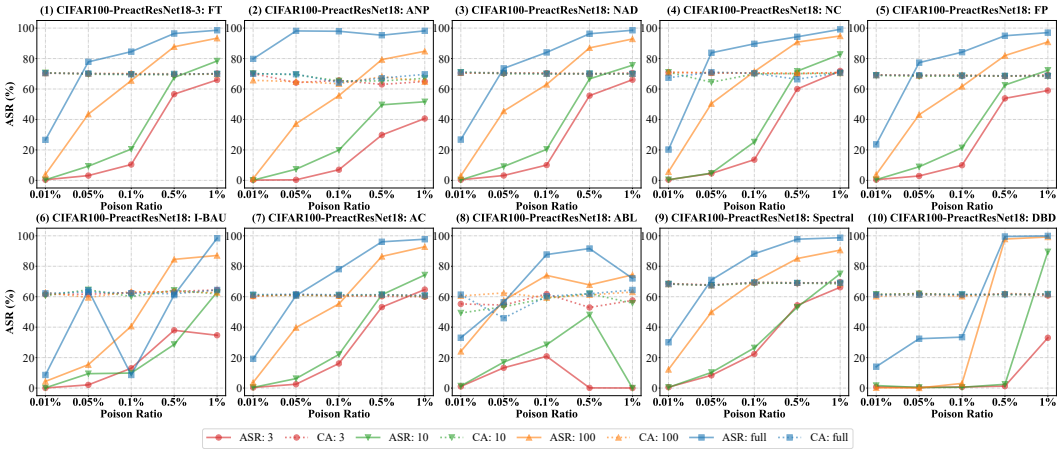


Figure 14: The defense results of SA and MA on CIFAR100 with PreactResNet18

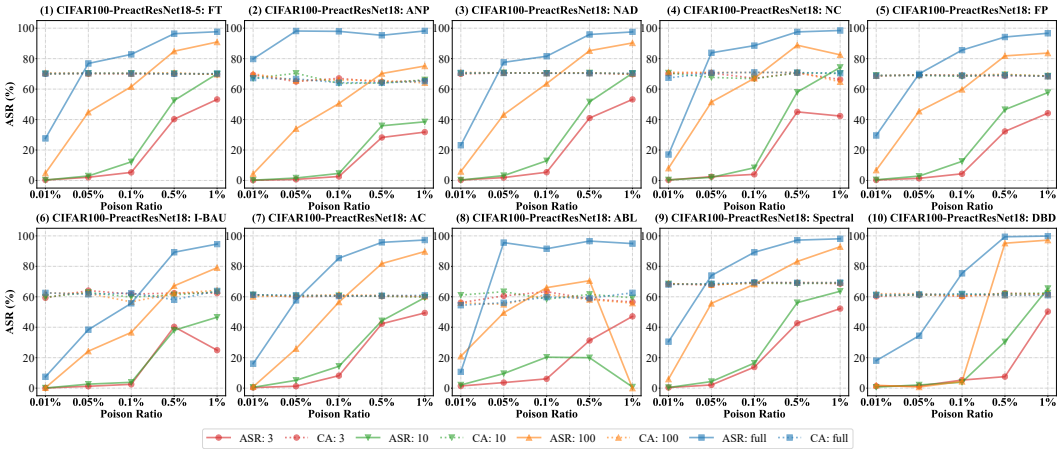


Figure 15: The defense results of SA and MA on CIFAR100 with PreactResNet18

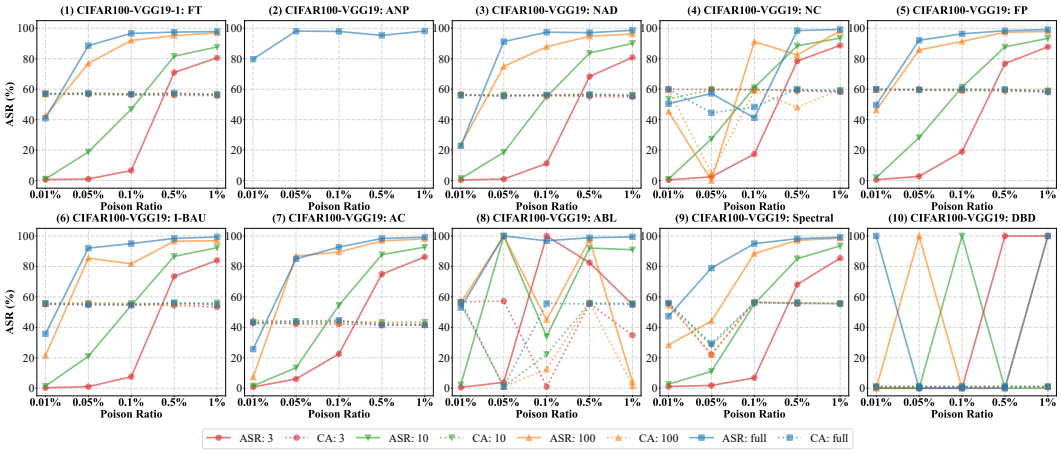


Figure 16: The defense results of SA and MA on CIFAR100 with VGG19

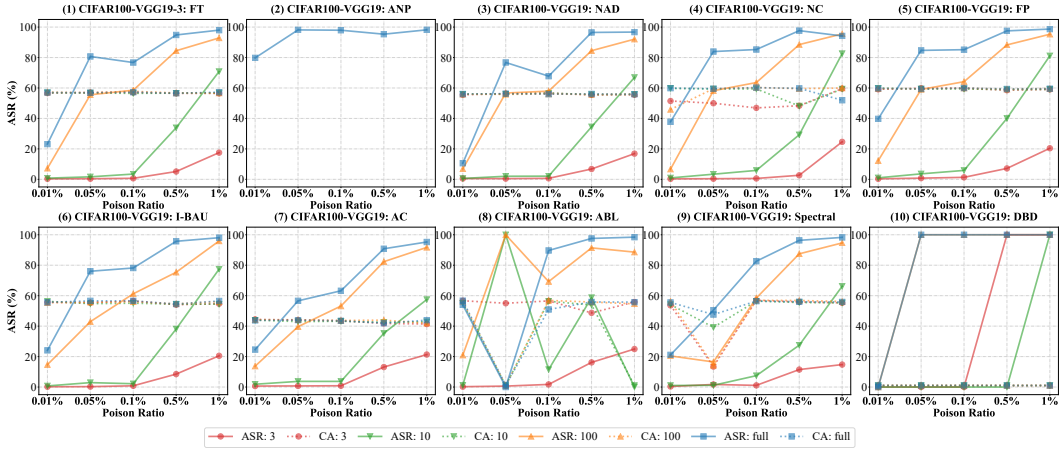


Figure 17: The defense results of SA and MA on CIFAR100 with VGG19

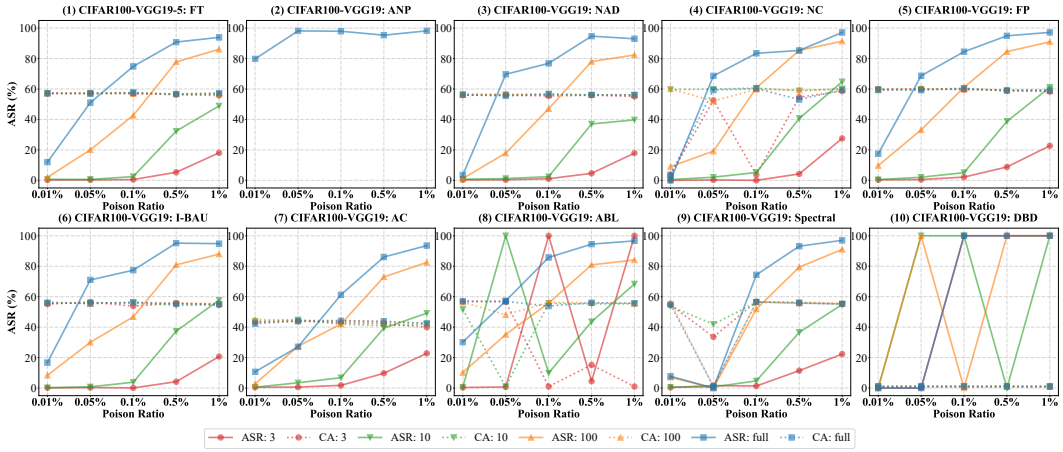


Figure 18: The defense results of SA and MA on CIFAR100 with VGG19

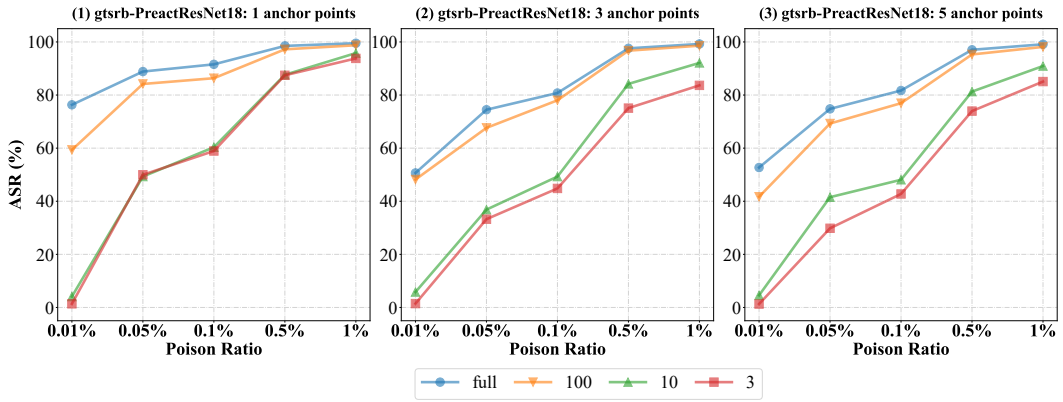


Figure 19: The attack results of SA and MA on GTSRB with PreactResNet18

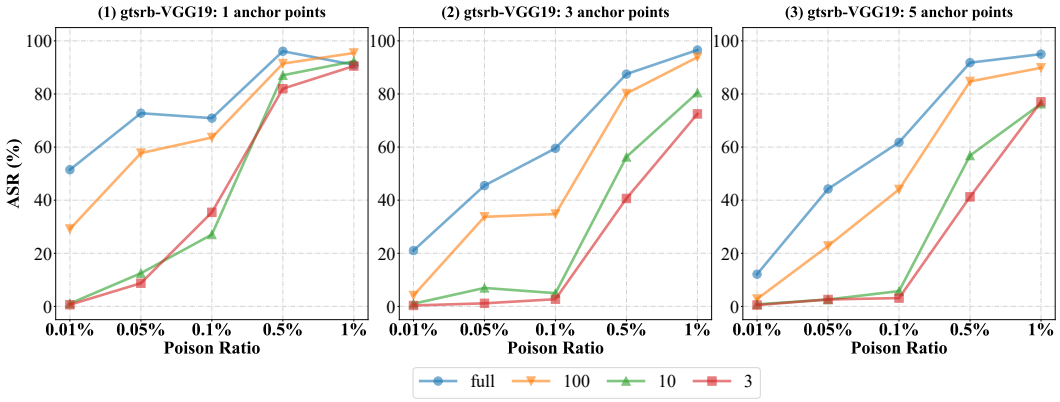


Figure 20: The attack results of SA and MA on GTSRB with VGG19

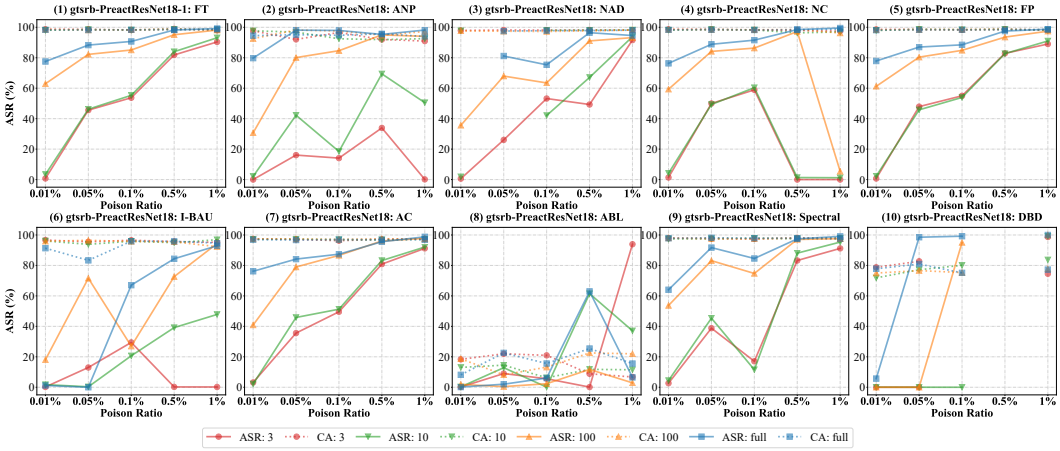


Figure 21: The defense results of SA and MA on GTSRB with PreactResNet18

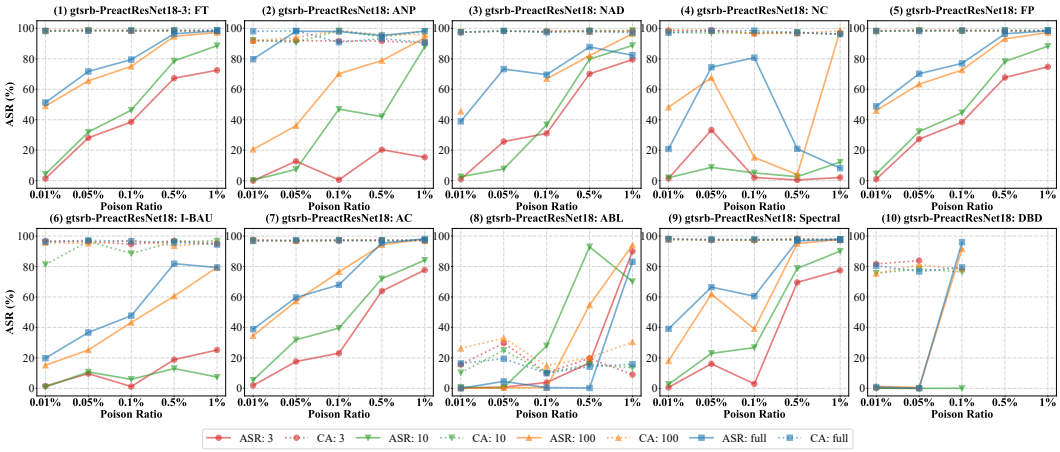


Figure 22: The defense results of SA and MA on GTSRB with PreactResNet18

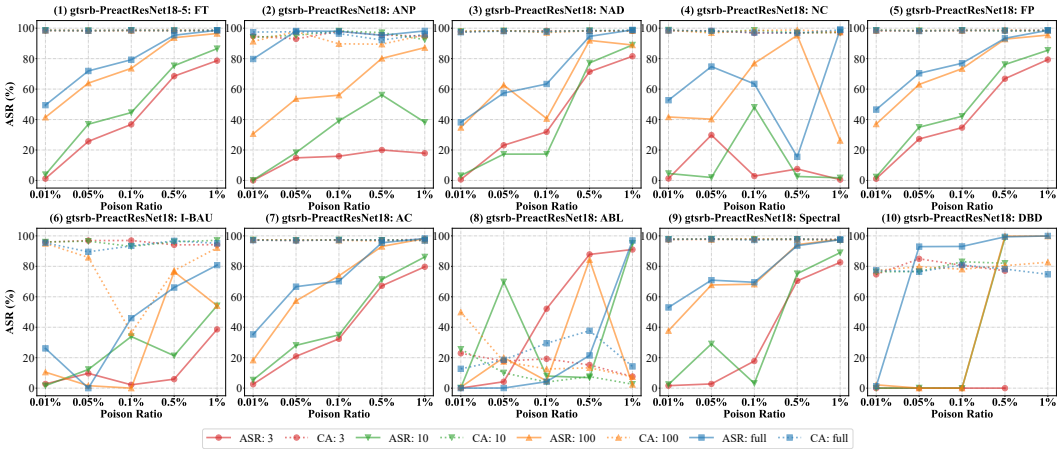


Figure 23: The defense results of SA and MA on GTSRB with PreactResNet18

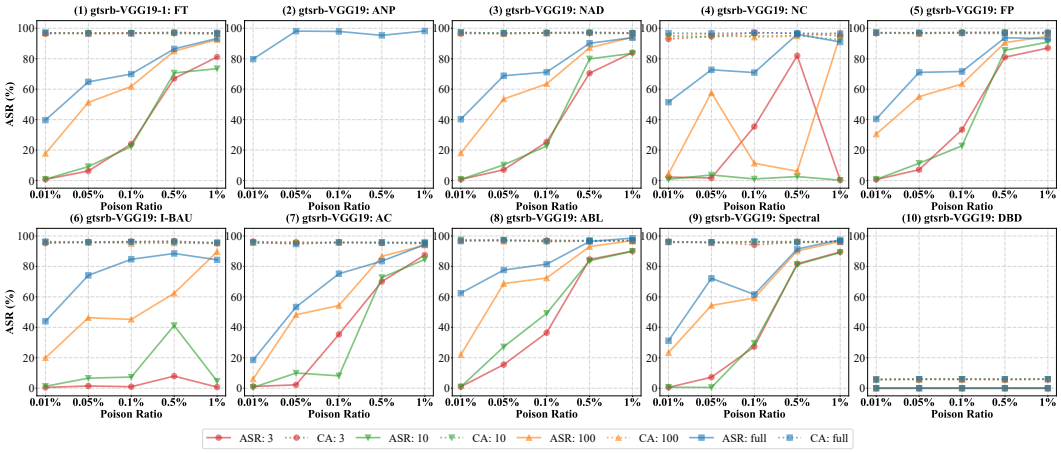


Figure 24: The defense results of SA and MA on GTSRB with VGG19

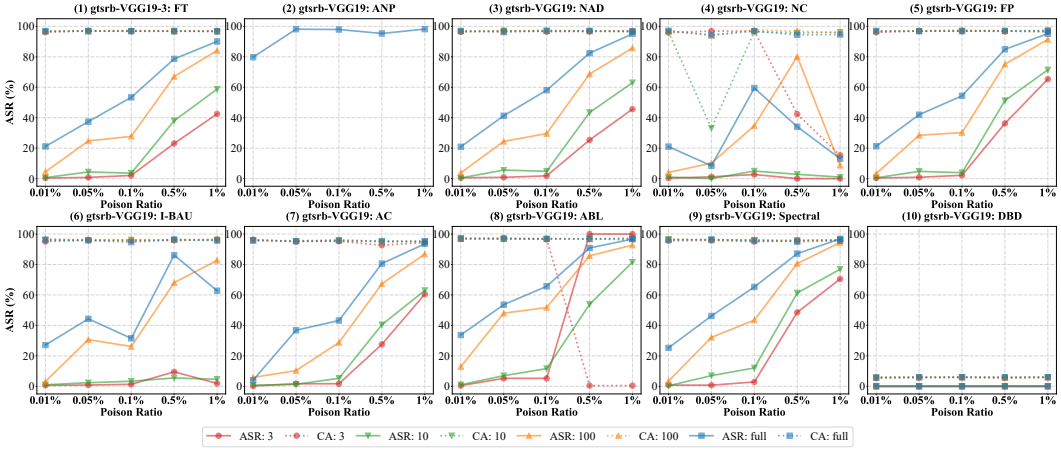


Figure 25: The defense results of SA and MA on GTSRB with VGG19

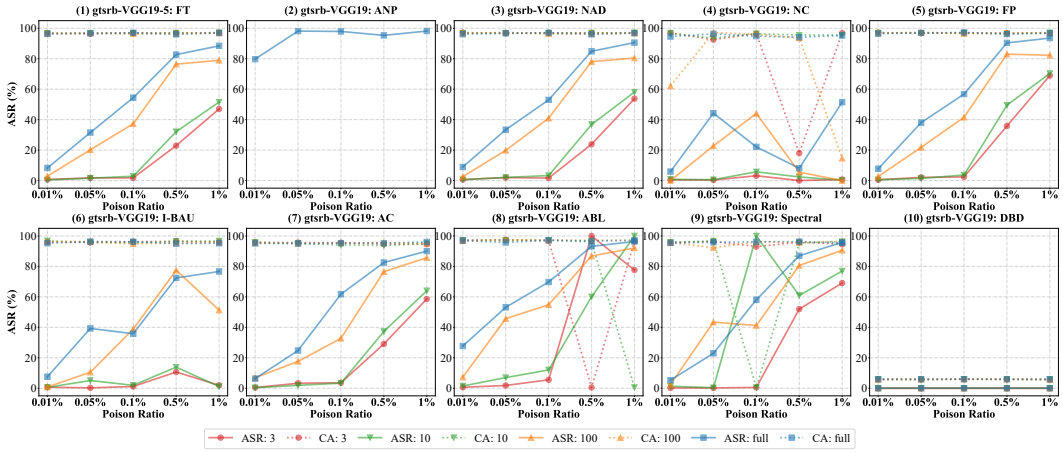


Figure 26: The defense results of SA and MA on GTSRB with VGG19