A Detailed Settings for backdoor attack

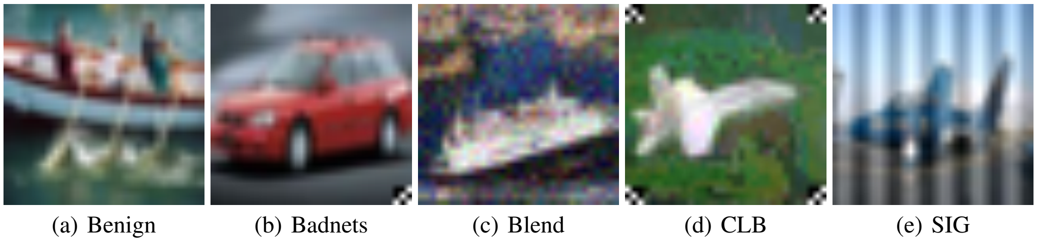

(a) Benign (b) Badnets (c) Blend (d) CLB (e) SIG

Figure 5: Examples for the benign and backdoor images in the poisoned training set.

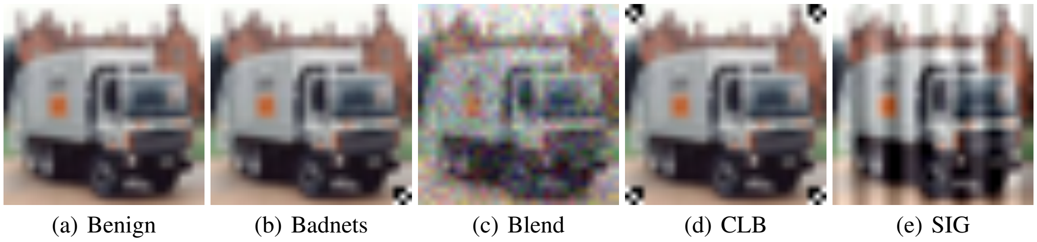

(a) Benign (b) Badnets (c) Blend (d) CLB (e) SIG

Figure 6: Examples for the benign and backdoor images in the poisoned test set.

This section provides detailed information about the settings for the backdoor attacks. As demonstrated in Section 3.1, we first pre-train the ViT-B on ImageNet-1k and finetune the network on the poisoned dataset using AdamW optimizer for 20 epochs with a learning rate of 0.0001. Simple data augmentations, including random crop with padding and horizontal flipping, are adopted for backdoor training. We assign the Class 0 (”airplane”) of the CIFAR-10 dataset as the target class for backdoor attacks. Examples of benign and backdoor images in the training set and poisoned test set are shown in Figure 5 and Figure 6. All experiments are performed on the NVIDIA 3090 GPUs. The implementation details of each attack are summarized as follows:

Badnets: Following the original paper (Gu et al., 2019), we take a 3×3 checkerboard as the trigger. As shown in Figure 5(b), the trigger is placed at the bottom right corner of the original image. Given the target class, 5% of images from the other classes are attached with the trigger and re-labeled as the target class. For ViT-B, we obtain the ACC of 97.85% and ASR of 100.00%.

Blend: For Blend attack, we take the Gaussian noise ($\boldsymbol{t}$) as the trigger. In particular, the trigger has the same size as the original image. For the benign image $\boldsymbol{x}$, the poisoned image can be given as $\boldsymbol{x}_p = (1-\alpha)\cdot\boldsymbol{x} + \alpha\cdot\boldsymbol{t}$. In contrast to the definition shown in Section 2.1, $\alpha \in [0,1]$ denotes as the blending rate between the benign image and the trigger pattern. Following the original paper (Chen et al., 2017), α is set to 0.2. Examples of poisoned images in the training and test set are shown in Figure 5(c) and Figure 6(c). Same as Badnets attack, 5% images from the other classes are attached with the trigger pattern and relabeled as Class 0. For ViT-B, we achieve the ACC of 97.85% and ASR of 100.00%.

CLB: We select 80% benign images from the target class for data poisoning. Next, we perform a 100-step PGD attack on the selected images using a pre-trained robust model [4]. For the hyperparameter settings, we follow the original paper with the budget $16/255$ and the step size of $2.4/255$. As shown in Figure 5(d), we attach the trigger, a four-corner 3×3 checkerboard, on these selected images. The poisoned training set combines these poisoned images and the remaining benign images from all classes. For ViT-B, we obtain the ACC of 97.83% and ASR of 96.23%.

SIG: We follow the original work in (Barni et al., 2019), which adopts the sinusoidal signal as the trigger. We also select 80% benign images from the target class for data poisoning. The strength Δ and frequency f for SIG attack are set to 40 and 6 respectively following previous studies (Wu et al., 2022; Barni et al., 2019). Examples of the poisoned images are shown in Figure 5(e) and Figure 6(e). For ViT-B, we obtain the ACC of 97.50% and ASR of 90.57%.

[4] https://github.com/yaircarmon/semisup-adv

Table 7: The effect of optimizer on FP and NAD. AdamW gains higher ACC and lower ASR than SGD.

(a) ACC

Attack		SGD		AdamW	
	No defense	FP	NAD	FP	NAD
Badnets	97.85	93.17	57.59	93.52	93.77
Blend	97.85	93.41	94.27	92.59	94.09
CLB	97.83	27.20	94.31	93.22	93.88
SIG	97.50	77.34	94.31	93.88	93.86
AvgDrop	-	24.98	12.91	4.46↓	3.86↓

(b) ASR

Attack		SGD		AdamW	
	No defense	FP	NAD	FP	NAD
Badnets	100.00	0.90	4.24	0.91	1.57
Blend	100.00	9.67	48.57	0.73	8.94
CLB	96.23	8.21	10.15	1.70	7.27
SIG	90.57	1.93	5.00	0.81	3.60
AvgDrop	-	91.53	79.71	95.66↑	91.36↑

B Detailed Settings for backdoor defense

This section provides detailed information on the backdoor defenses applied in this paper. The settings of each defense are summarized as follows:

FT: We use AdamW (Loshchilov & Hutter, 2018) optimizer, the most popular optimizer for ViTs, to fine-tune the backdoor ViTs for 20 epochs with a learning rate of 3e-4 and a weight decay of 0.2. In addition, we adopt the cosine learning rate schedule. Same as backdoor training, only simple data augmentations, including random crop with padding and horizontal flipping, are used to retain the clean accuracy better and avoid the increasing ASR of whole-image backdoor attacks caused by strong data augmentation as discussed in section 3.

FP: FP (Liu et al., 2018a) first prunes the last layer of CNNs by a predefined pruning threshold and then fine-tune the network on the clean subset of data. Similarly, we prune the last linear projection layer of transformer encoder blocks in ViTs. For the pruning partition threshold, we use *the tolerance of clean accuracy reduction* to limit the maximum drop of the benign accuracy following (Wu et al., 2022). In this paper, we set it to 0.9. The other settings are the same as the original paper (Liu et al., 2018a).

NAD: NAD (Li et al., 2021) first makes two copies of the original backdoor models, referred to as the teacher model and student model respectively. Next, NAD fine-tunes the teacher model with the vanilla FT. Finally, the finetuning of the student model is guided through neural attention transfer from the teacher model. For the hyperparameter setting, we mainly keep in line with (Wu et al., 2022) except for two differences: we train the student network for 20 epochs using the AdamW optimizer instead of hundreds of epochs with SGD optimizer. The above changes are made because of the observation shown in Appendix C and Appendix D.

ANP: Wu et al. (Wu et al., 2020) observe that backdoor models are prone to output the target labels when the neurons are perturbed by the adversarial perturbations. Inspired by this, they propose to optimize the mask of each neuron, a continuous value in $[0, 1]$, under adversarial neuron perturbations and then prune neurons whose mask values are lower than the threshold, *i.e.*, hardening the continuous mask values as binary masks. In this paper, we use the same settings as the original paper except for applying 4000 iterations to avoid under-convergence of large models like ViTs (longer than the 2000 iterations for CNNs in the original paper). Compared to the hardened masks (pruned) applied in their original paper, we find that soft masks, continuous mask values without hardening, can preserve ACC better and decrease ASR further. Thus, we apply soft masks in this paper, and these masks are applied to the channels of linear projection.

AWM: Compared to ANP, AWM (Chai & Chen, 2022) makes two improvements on CNNs. The authors apply soft element-wise weight masking instead of neuron pruning (hardened mask values) to avoid over-cutting beneficial information. Besides, they perturb the data instead of the neurons to utilize the training data more efficiently. When applied to ViTs, we mask the channel of the linear projection, similar to ANP. The other hyperparameters are the same as the original paper (Chai & Chen, 2022) without turning.

C The Effect of Optimizer on FP and NAD

In this section, we compare the performance of SGD and AdamW on the other two fine-tuning-based methods, FP and NAD, following the settings in section 3.2. As shown in Table 7, the results demonstrate that, compared to SGD, AdamW always performs better on FP and NAD. For example,

Table 9: ACC (%) of our attacks with different ViT variants on the benchmark dataset. The best results are in **bold**.

Defense	Attack	Vanilla					Ours				
		ViT-B	DeiT-S	Swin-B	Cait-S	XciT-S	ViT-B	DeiT-S	Swin-B	Cait-S	XciT-S
No defense	BadNets	97.85	97.67	98.53	98.47	97.83	**98.18**	**97.75**	**98.69**	98.35	**97.90**
	Blend	97.85	**97.98**	**98.90**	**98.62**	**98.39**	**98.04**	97.86	98.75	98.47	98.34
	CLB	97.83	97.70	98.41	98.27	97.65	**97.88**	**97.83**	**98.49**	98.27	**97.72**
	SIG	97.50	**97.44**	98.56	**98.21**	**98.05**	**97.88**	97.36	**98.67**	98.14	97.89
FT	BadNets	93.79	**94.29**	96.64	96.09	**95.82**	**94.03**	94.16	**96.86**	**96.66**	95.52
	Blend	93.30	**94.07**	95.96	**96.83**	**96.06**	**94.00**	93.99	**96.83**	96.59	95.89
	CLB	94.06	**94.28**	**96.67**	96.39	95.53	**94.20**	94.01	96.24	**96.50**	**95.92**
	SIG	**93.51**	**93.98**	96.78	96.52	95.84	93.45	93.79	**97.14**	**96.59**	**95.96**
FP	BadNets	93.52	93.40	95.84	95.18	**94.57**	**93.67**	**93.41**	**95.98**	**95.29**	93.59
	Blend	92.59	**94.06**	**95.94**	**94.69**	94.37	**93.05**	93.96	96.11	95.43	**94.79**
	CLB	**93.22**	93.99	95.91	95.36	**94.55**	93.15	**94.17**	**95.48**	**95.42**	94.36
	SIG	**93.88**	93.36	95.97	**95.50**	**94.54**	93.75	**93.84**	**96.24**	95.20	94.37
NAD	BadNets	93.77	**95.39**	97.03	**97.00**	95.76	**93.82**	95.19	**97.12**	96.91	**95.85**
	Blend	94.09	**95.85**	**97.12**	**96.77**	**95.93**	**94.12**	95.57	97.08	96.51	95.92
	CLB	93.88	**95.38**	**96.89**	**96.98**	95.87	94.02	95.09	96.75	96.57	**96.52**
	SIG	93.86	**95.51**	97.20	96.95	**96.23**	**93.95**	95.22	**97.52**	96.95	95.62
ANP	BadNets	94.26	95.86	**98.18**	**97.59**	**97.14**	**94.40**	**96.26**	98.12	97.56	96.68
	Blend	92.70	96.47	**98.18**	98.00	**97.14**	**95.67**	**96.70**	98.14	**98.47**	96.68
	CLB	95.71	96.45	97.89	97.61	**97.33**	**95.83**	**96.68**	**98.12**	**97.71**	96.97
	SIG	92.60	96.55	97.87	**97.73**	**97.91**	**94.62**	96.55	**98.01**	97.69	97.47
AWM	BadNets	**95.02**	94.52	**96.39**	95.93	**95.46**	93.87	**94.91**	96.28	**96.18**	95.43
	Blend	**95.08**	**94.99**	93.00	**96.51**	**96.00**	95.06	94.82	**95.38**	96.28	94.40
	CLB	**95.60**	**94.94**	**95.20**	96.17	95.33	95.12	94.84	94.22	**96.41**	**95.53**
	SIG	**94.58**	**94.76**	96.89	**96.59**	**96.05**	94.46	94.43	**96.90**	96.57	95.80

SGD results in an average ACC drop of 24% in FP, much larger than 4.46% caused by AdamW. Besides, SGD also has a little worse defense performance.

D THE EFFECT OF FINE-TUNING EPOCHS ON FT, FP AND NAD

Table 8: The performance of Fine-tuning-based defenses for different fine-tuning epochs.

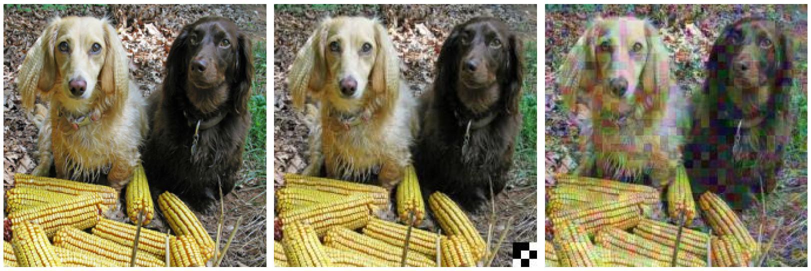
Metric	Defense	epoch=20				epoch=100				
		Badnets	Blend	CLB	SIG	Badnets	Blend	CLB	SIG	AvgDrop
ACC	FT	93.79	93.30	94.06	93.51	90.30	90.43	91.20	90.19	3.14
	FP	93.52	92.59	93.22	93.88	89.86	90.01	89.56	89.45	3.58
	NAD	93.77	94.09	93.88	93.86	90.62	91.22	90.87	91.14	2.94
ASR	FT	2.51	4.91	1.33	1.40	1.26	3.15	1.48	0.93	0.83
	FP	0.91	0.73	1.70	0.81	1.08	1.01	2.13	0.80	-0.22
	NAD	1.57	8.94	7.27	3.60	1.49	4.62	5.08	2.59	1.89

Here, we compare the performance of the fine-tuning-based methods for different fine-tuning epochs. As shown in Table 8, a notable accuracy drop appears on all defenses when we fine-tune the models for longer epochs, *e.g.*, the average accuracy drop is 3.14% in FT, which hinders the use of the model. With such a notable accuracy drop, ASR only decreases slightly, *e.g.*, 0.83% in FT with more epochs. Therefore, we recommend using fewer epochs to preserve the utility of the ViTs better.

E THE ACCURACY OF OUR ATTACK ON CIFAR-10 DATASET

We have discussed the attack performance of our proposed method as shown in Table 4 of Section 5.1. Here, we continue to explore the effect on the accuracy of our attacks. As shown in Table 9, the backdoored models with our method have comparable accuracy to their baselines (without our method), which indicates our method does not influence the utility of the backdoored model and guarantees the stealthiness of backdoored models with our method.

F The Setting of Our Attack on ImageNet Dataset



(a) Benign (b) Badnets (c) Blend

Figure 7: Examples for the benign and backdoor images on ImageNet dataset.

Attack: Since the huge computational cost, we fine-tune the pre-trained ViT-B on the poisoned ImageNet with 512 batch size and 10 epochs to insert backdoors. Because ImageNet is a high-resolution dataset, we increase the trigger size of badnets attacks to 21×21 for better poisoning. For the Blend attack, we resize the image of gaussian noise to 224×224 to accommodate the large input size on ImageNet. In Figure 7, we show examples of benign and backdoor images. For other settings of the vanilla poisoning, we keep the same with our experiments on CIFAR-10 (Please refer to Appendix A for details.). For the settings of our proposed attack, we follow the settings of CIFAR-10 except for the following two points: During the perturbation generation step, the budget and step size are set to $8/255$ and $2/255$, respectively. Similar to the vanilla backdoor attack, the patch size of RMP is enlarged to 16 because ImageNet is a high-resolution dataset. For ViT-specific attacks, we choose DeiT-B (Touvron et al., 2022) which has the exact same architecture as ViT-B for poisoning without any hyperparameter change.

Defense: First, for the defense methods unrelated to architectures, to achieve a better acceleration of the experiments on ImageNet, we adopt a large batch size of images for defense. In detail, for fine-tuning-based defense, the batch size is set to 512. For pruning-based defense, the batch size is set to 128 to avoid the out-of-memory problem on 4 NVIDIA 3090 GPUs. Other settings are the same as our experiment on CIFAR-10. Please refer to Appendix B for details. As for the ViT-specific attack: attention blocking (AB), we adopt the default setting recommended by (Subramanya et al., 2022b): during the inference stage, we block out the region of size 30×30 which is highlighted by Attention Rollout (Abnar & Zuidema, 2020).

G The Accuracy of Our Attack on ImageNet Dataset

Like the experiments on CIFAR-10, we also evaluate the effect of our method on ACC for large datasets like ImageNet. The results in Table 10 show that our method does not influence the utility of the backdoored models and the stealthiness of backdoored models on large datasets can also be further guaranteed.

Table 10: ACC (%) of our attack on ImageNet dataset. The higher ACC is in **bold**.

Attack	Before	FT	FP	NAD	ANP	AWM	AB
TrojViT	80.59	76.82	76.93	77.55	76.31	77.78	-
DBIA	79.52	78.3	75.2	77.18	76.49	78.94	-
Badnets	80.82	71.05	68.10	72.38	69.56	76.40	**74.86**
CAT+Badnets	81.01	**71.41**	**68.31**	**72.69**	**69.79**	**76.62**	74.51
Blend	80.82	71.03	**68.43**	72.60	69.69	**76.77**	74.72
CAT+Blend	**81.03**	**71.12**	68.39	**72.62**	**69.96**	76.36	**74.73**

H Broader Impact

While our adaptation to backdoor defense eliminates backdoor behaviors inside backdoored ViTs, it is important to avoid creating overconfidence among readers regarding the robustness of current ViTs.

Note that there still may exist powerful attacks that can bypass these existing defenses, like the new attack we proposed in this paper. Furthermore, the proposed method is a strong attack to existing defense, thereby increasing potential risks in practical applications. However, we firmly believe that comprehensive evaluations using stronger attacks and more revealed potential risks would encourage practitioners to prioritize the security of their deployed models.