Rethink Backdoor Robustness in Vision Transformers

Yichuan Mo[1], Dongxian Wu[2], Yifei Wang[3⋆], and Yisen Wang[†]

[1] State Key Lab of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University, China
mo666666@stu.pku.edu.cn, yisen.wang@pku.edu.cn
[2] Independent Researcher, China
wudx16@gmail.com
[3] Amazon AGI SF Lab, USA
yifeiwg@amazon.com

Abstract. Backdoor attacks, which induce Convolutional Neural Networks (CNNs) to behave maliciously when a predefined trigger is present, pose serious security risks. While such threats also extend to Vision Transformers (ViTs), previous studies have suggested that existing backdoor attacks remain highly effective on ViTs and can evade common defense mechanisms—often without significantly compromising accuracy. In this paper, we revisit this claim and demonstrate that such conclusions are overly optimistic, largely due to inadequate adaptation of CNN-based defenses to ViTs. We show that, with proper adjustments, existing backdoor attacks can in fact be effectively mitigated. Moreover, we propose a more robust attack strategy: by introducing slight perturbations to the trigger, existing attacks can be made significantly more resistant to various defenses. We hope that our findings—both on the defensibility of current attacks with correct adaptations and the proposed enhanced attack—will inspire deeper investigation into the backdoor robustness of Vision Transformers. Our code is available at https://github.com/PKU-ML/ViT_backdoor.

Keywords: Vision Transformer · Backdoor Attack · Backdoor Defense

1 Introduction

Vision Transformers (ViTs) [14, 26] have demonstrated outstanding performance in various tasks, including image classification [43, 55], semantic segmentation [39], and image generation [5, 18], leading to their widespread popularity. However, strong performance alone is insufficient for ViT to be practically deployable. It must also exhibit trustworthiness without posing severe security risks, and one of the most notable ones is the backdoor attacks [9, 16], which implant unexpected behaviors inside models, making the victim model

⋆ The work was done at MIT prior to joining Amazon.
† Corresponding Author.

produce specific misclassification in the presence of a predefined trigger while maintaining high performance on benign images. While previous studies mainly focus on Convolutional Neural Networks (CNNs), there is a growing need for an in-depth investigation of ViTs to help practitioners better understand the potential risks and deploy them more reliably.

Table 1: The performance of FT against Badnets attack for ResNet-18 and ViT-B on CIFAR-10 [51].

	ResNet18	ViT-B
ASR	1.48%	8.81%
ACC	89.96%	42.00%

After a long arms race between backdoor attack and defense, even relatively simple defenses can make attacks fail on CNNs. For example, under fine-tuning defense against Badnets in Table 1, ResNet18 achieves only 1.48% attack success rate (ASR) while maintaining 89.96% benign accuracy (ACC), indicating attack failure. In contrast, ViTs under the same setting show higher ASR and lower ACC, suggesting disrupted benign utility. Since Badnets is model-agnostic, this disparity motivates us to investigate its underlying cause.

Drawing inspiration from [31], we distinguish a crucial observation: 1) CNNs are usually trained by SGD and its fine-tuning defense is also trained by SGD; 2) ViTs are typically trained by AdamW [27] while its fine-tuning defense is trained by SGD (NOT AdamW, inheriting from earliest work [14], which first introduces transformers to computer vision). This discrepancy in optimizers raises the possibility that the perceived vulnerability of ViTs (with defense) might be overstated. In this paper, we first conduct a series of experiments to comprehensively investigate the above hypothesis, which is further confirmed that the threat posed to ViTs with defense has been magnified. Upon minor modifications, ViTs with existing backdoor defense methods demonstrate clear resistance to attacks, mirroring the robustness of CNNs.

To this end, we are wondering whether a more powerful attack exists that can better evade current defenses. Therefore, we analyze backdoored models and further propose a simple yet effective attack. We discover that it is easy for backdoor defenses to detect and utilize the differences in channel activations due to the noticeable difference in the intermediate layers between the inputs with and without triggers. However, we can reduce this difference by adding small perturbations to the triggers before training while keeping triggers unchanged during inference, resulting in more reliable backdoor attacks. Additionally, our method has transferability across different transformer architectures and is effective for both small and large datasets.

In summary, our contributions are summarized as follows:

- We investigate the existing backdoor defenses on ViTs and find the outstanding performance of the backdoor attacks to ViTs is over-estimated due to the inappropriate adaptations. Further, we provide a practical training recipe to improve the performance of defense and show that existing attacks on ViTs can still be easily resisted.
- We propose to add small perturbations to the triggers before training to suppress the difference in the intermediate-level representations between the

inputs with and without triggers, proposing a more powerful attack, which is termed the channel activation attack in ViTs (CAT).
– Our contributions, including the finding of existing attacks to current defenses and the development of a new attack, contribute to a reliable baseline for the backdoor robustness of ViTs. We hope it can be a cornerstone of future studies in improving the backdoor robustness of ViTs.

2 Related Work

2.1 Backdoor Attack

Backdoor attacks [9, 16, 38, 53], also known as Trojan attacks, indicate the behaviors of implanting specific malicious behavior into machine learning models, which make the models perform well on benign data while leading to specific misclassifications on inputs containing triggers (*i.e.*, triggered inputs). The adversary usually poisons the training data [59] or controls the training process [25] to achieve this. Typically, a trigger pattern is added to the input image as follows,

$$\mathbf{x}_p = (\mathbf{1} - \mathbf{m}) \odot \mathbf{x} + \mathbf{m} \odot \mathbf{t}, \tag{1}$$

where $\mathbf{t}$ is the trigger pattern and mask $\mathbf{m}$ indicates the pixels affected by the trigger pattern. Usually, the adversary re-labels the triggered input as the predefined target class (*i.e.* in a dirty-label setting). Models trained on a mixture of these poisoned data and other benign data are implanted with an unexpected correlation between the trigger pattern and the target class. To improve the attack stealthiness, some studies explored less noticeable trigger designs like the semi-transparent ones [9], the elastic transformed ones [33], and the input-aware ones [32]. Besides, since incorrect annotation might expose the existence of the poisoned data, some studies focus on attacks without re-labeling (clean-label settings) [6,36, 46, 54, 58]. Although most previous backdoor attacks focus on CNNs, researchers have started to focus on backdoor attacks on ViT regarding their popularity. Although ViTs are reported to be more robust against adversarial attacks [2, 37] and common corruption [4, 7], they are still vulnerable to backdoor attacks [28, 41]. Reliable attacks are needed to help practitioners properly understand the risks of backdoor threats and deploy these models reliably.

2.2 Backdoor Defense

To mitigate the potential risks caused by backdoor attacks, numerous studies proposed various defense methods, mainly categorized into **defense during training** and **defense after training**. Defense during training attempts to mitigate the impact of poisoned data in the training set. Some methods detect and remove poisoned data by treating them as outliers [10, 15, 47], some employ semi-supervised learning to bypass the incorrect correlations [19], and others utilize differential privacy to ensure that a poisoned portion of training data is unable to cause severe results [30]. Meanwhile, the defense after training [62]

includes those that detect the backdoor samples from the model inputs [15, 45] or purify the model to removes the backdoor behavior inside DNNs. Since the latter category is closer to the ideal goal, it has received more attention. This can be accomplished by fine-tuning the model using a small amount of clean data [35] (FT) and further improving it with neuron pruning [24] or attention alignment [22]. Since the performances of fine-tuning are easy to suffer a substantial decrease when the data is limited, another popular method is selectively removing neurons related to the backdoor behaviors [8, 48, 52]: Built upon the observation that the backdoor behavior can be revealed by the adversarial neuron perturbation, ANP [52] formulates the following min-max problem with dataset D_v to expose the malicious neuron:

$$\begin{aligned} &\min_{\mathbf{m}\in[0,1]^n} \big[\alpha \mathcal{L}_{D_v}(\mathbf{m}\odot\mathbf{w},\mathbf{b}) \\ &+(1-\alpha)\max_{\delta,\xi\in[-\epsilon,\epsilon]^n} \mathcal{L}_{D_v}((\mathbf{m}+\delta)\odot\mathbf{w},(1+\xi)\mathbf{b})\big], \end{aligned} \tag{2}$$

where δ and ξ are the perturbations to the weight $\mathbf{w}$ and bias $\mathbf{b}$ of all neurons respectively. They maximize the cross-entropy loss $\mathcal{L}_{D_v}$ and $\mathbf{m}$ is the mask that adversarially preserves the clean accuracy and covers up the backdoor behavior. Then the neurons corresponding to low mask values are pruned to purify the backdoor model. As an improved approach based on ANP, AWM in [8] proposes to adopt the element-wise weight and perturb the input data instead to gain better performances on small networks. This paper primarily focuses on defense after training. Because ViTs demand a large amount of data and extensive training resources, it has become impractical for most practitioners to train ViTs from scratch, making defense after training a more realistic scenario. Previous studies [51, 56] suggested that directly applying defenses from CNNs to ViTs fails. For example, fine-tuning decreases natural accuracy from 94.58% to 42.00% against the Badnets attack and fine-pruning totally collapses in [56]. At the meantime, only a few defense methods specially designed for ViT are proposed [13, 40] and their performance is lagging far behind the state-of-the-art defense on CNNs: The adaptive defense proposed in [61] only decreases the ASR of TrojViT (a ViT-specific attack) to 77.13% and the patch processing method in [13] fails to detect 33.2% backdoor examples on CIFAR-10. It seems that existing attacks can already obtain outstanding performances on resisting defense for ViTs. However, in this paper, after re-investigating various backdoor defenses with ViTs, we reveal that the achievement obtained by previous attacks is attributed to the improper adaptation of defenses. Furthermore, we provide a new attack, based on the empirical observation of the channel activations to help existing attacks evade defenses more effectively.

3 The Vulnerability of Defense on ViTs to Existing Attacks

In this section, we reevaluate the perceived susceptibility of ViTs to prevailing backdoor attacks when equipped with potential defenses. We primarily consider

two categories of defenses: one is fine-tuning-based, including Fine-Tuning (FT), Fine-Pruning (FP) [24], and Neural Attention Distillation (NAD) [22], Fine-tuning with SAM optimizer (FT-SAM) [63], Super Fine-tuning (Super-FT) [35] and the other is pruning-based, including Adversarial Neuron Pruning (ANP) [52] and Adversarial Weight Masking (AWM) [8].

3.1 Threat Model and Basic Settings

Threat Model: Consistent with the assumptions made by most attacks [51], our threat model limits the capability of attackers to the access of training data. An unaware third party trains a ViT model using the poisoned data and helps downstream users who lack the training resources to solve the downstream tasks. The defenders provides service for the users and ensure that the downloaded parameters from the third party are free of backdoor threats.

Settings: We train a backdoored ViT-B [14] with various attack methods. Specifically, we initialize the model with a pre-trained weight [50] on the ImageNet-1k [12] and then fine-tune it on CIFAR-10[4] [21]. Note that a portion of CIFAR-10 training data is poisoned to implant the backdoor behavior, *i.e.*, some images are added with the trigger pattern and are re-labeled as the target class if expected. We apply six commonly-used attack methods: 1) Badnets [17], 2) Blend [9], 3) CLB [46], 4) SIG [6], 5) IAD [32], 6) SSBA [23]. Their trigger design and poisoning method in the original paper are kept. For dirty-label attacks, we set the poison rate as 5% and for clean-label attacks, we poisoned 80% images of the target class. To accommodate the input size of ViT, we first add triggers to CIFAR-10 images (32×32) and then resize them to a larger size (224×224). For more detailed information, please refer to Appendix A. Here, we use accuracy (ACC) to indicate the classification performance on benign data, and attack success rate (ASR), the percentage of triggered input being classified as the target class, to indicate the attack performance. Note that we will remove the inputs whose ground-truth label is the target class, and thus, a successful defense should make ASR as low as 0.

3.2 ViTs with Fine-tuning-based Defense

Fine-tuning is one of the most basic and model-agnostic defenses. However, as discussed in Section 1, directly inheriting fine-tuning-based defense strategies from CNNs can potentially lead to suboptimal outcomes. Note that SGD is the commonly used optimizer for both training and fine-tuning for CNNs, while for ViTs, the first work [14] introducing Transformers to computer vision, adopts AdamW for pre-training and SGD for fine-tuning. Notably, prior work [51] on backdoor defense naturally inherit this strategy and observes notably diminished accuracy across multiple backdoor attacks. This discrepancy in optimizers motivates us to study the potential influence of optimizers on backdoor defense. The

[4] Only 95% of the original training data on CIFAR-10 are used to train the backdoored model, and the remaining data are kept for defense.

Table 2: The comparison between SGD and AdamW optimizer on FT. Here, AvgDrop represents the average drop of six attacks on ASR/ACC after performing FT.

Attack	ACC			ASR		
	No defense	SGD	AdamW	No defense	SGD	AdamW
Badnets	97.85	14.32	95.14	100.00	5.02	0.72
Blend	97.85	10.99	95.32	100.00	7.10	5.22
CLB	97.83	15.67	95.15	96.23	6.80	3.39
SIG	97.50	12.62	95.32	90.57	12.12	5.66
IAD	97.79	15.00	95.57	100.00	13.80	5.36
SSBA	98.19	11.78	96.05	99.23	9.16	0.50
AvgDrop	-	84.44	**2.41**	-	88.67	**94.20**

initial learning rates for SGD and AdamW are set to 0.02 and 3e-4, respectively. For the other parameters in AdamW, including data augmentations such as Mixup [60] and CutMix [57], we use the settings summarized in Appendix B. Table 2 illustrates the experimental fine-tuning (FT) results against various backdoor attacks. For the results on other fine-tuning-based attacks, please refer to Appendix C. We find that SGD exhibited significant instability on ViTs: Its ACC is below 20% under all attacks. In contrast, AdamW can achieves not only high ACC and but also low ASR under all settings. Therefore, simply using SGD for backdoor defense on ViTs will yield highly unstable performance. In Appendix D, we analyze the reason behind it from an empirical view.

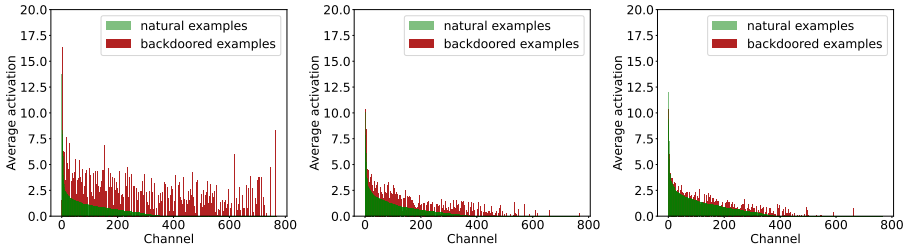
3.3 ViTs with Pruning-based Defense

Pruning is also a typical defense approach, which attempts to remove backdoor-related neurons/channels and is severely impacted by the architectures. In previous studies, pruning-based methods have achieved excellent robustness against backdoor attacks with CNNs [8, 52]. However, when we directly apply these methods to ViTs, we find that they are unable to effectively defend as shown in Table 3. Specifically, ANP fails to reduce ASR and cannot remove the backdoor-related neurons. Besides, although AWM reduces ASR, it also largely decreases ACC, making the model unusable. To explore the potential reason, we look deeply at the implementation of ANP and find that ANP actually prunes channels inside norm layers rather than neurons inside convolutional layers. This is because, in CNNs, each neuron is typically surrounded by at least one norm layer[5]. However, in ViT, many norm layers are removed, and norm-layer-based pruning only influences part of neurons and limits the defense performance. Meanwhile, AWM utilizes element-wise masks for optimization, whose number of parameters is the same as the total number of parameters of ViT. Since ViTs are typically larger and lack inductive bias, AWM encounters the severe overfitting issue, leading to

[5] Specifically, for Preact-ResNet, the norm layer is always located before the neuron; for ResNet, it is located after the neuron

Table 3: The Performance of Pruning-based Defense with or without adaptation. Here, AvgDrop represents the average drop of six attacks on ASR/ACC after performing the defense.

Attack	No defense		ANP		ANP Adapted		AWM		AWM Adapted	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Badnets	97.85	100.00	97.85	100.00	94.26	1.34	85.98	1.24	95.02	0.71
Blend	97.85	100.00	97.85	100.00	92.70	23.70	83.29	2.03	95.08	1.70
CLB	97.83	96.23	97.83	96.23	95.71	12.71	85.67	3.48	95.60	1.52
SIG	97.50	90.57	97.50	90.57	92.60	1.48	87.22	1.16	94.58	3.87
SSBA	98.19	99.23	98.19	99.23	93.88	0.23	82.60	1.53	95.64	1.24
IAD	97.79	100.00	97.79	100.00	92.91	5.64	88.07	6.49	94.47	6.38
AvgDrop	-	-	**0.00**	0.00	4.19	**90.16**	12.36	**95.17**	**2.77**	95.10



(a) Before defense **(b)** After FT **(c)** After AWM

Fig. 1: The average activations for different channels before (a) and after the backdoor defense (b)-(c). The activations are sorted in descending order of the activations on natural samples.

low accuracy. Therefore, to make pruning methods applicable to ViTs, selecting appropriate granularity and pruning locations is necessary. Here, we recommend directly pruning all channels of linear projection inside both attention and MLP layers, which provides better coverage than ANP and requires fewer parameters compared to AWM. As shown in Table 3, this modification decreases ASR notably and keeps ACC high.

4 Proposed Backdoor Attack – CAT

Following the above analysis, existing defense methods (ViTs adapted) successfully defend against backdoor attacks in ViTs, just as they do in CNNs. Here, we want to explore whether there exist new backdoor attacks to beat the newly adapted defense on ViTs.

To obtain a better insight into why defense methods can detect and remove backdoor behaviors, we investigate the per-channel activations before the MLP head in ViT. We illustrate the average activations of all channels for a backdoored ViT-B on triggered and benign inputs from the CIFAR-10 test set, respectively. For

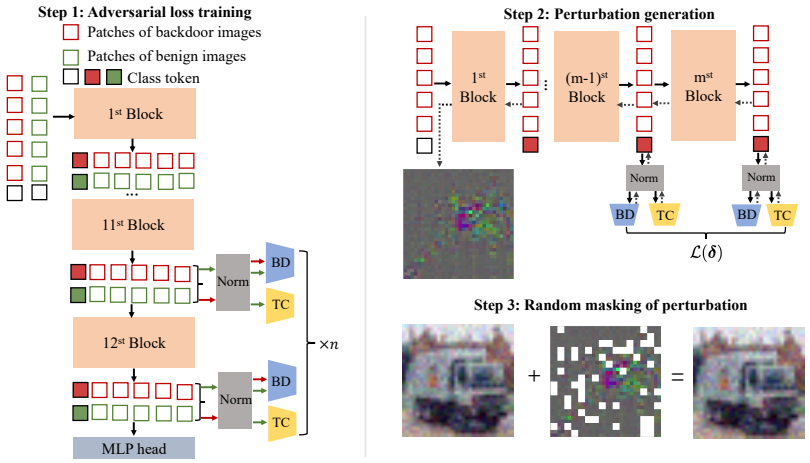


Fig. 2: The illustration of our proposed attack. We illustrate our attack by taking ViT-B as an example. *left:* Using the existing poisoned dataset, we firstly train the BD and TC on a surrogate model (**Step 1**). *right:* When the training is over, we perform adversarial attacks on the BD and TC modules to generate adversarial perturbation (**Step 2**). In each step during crafting adversarial perturbations, we manually mask some patches of perturbation to better poison ViTs (**Step 3**).

clarity, we reorganize the channels based on their average activations, arranging them from largest to smallest with respect to average activations on benign data. In Figure 1, we find a significant activation difference between benign and triggered inputs, which is easy to capture. Further, we compare the average activation of all channels for models purified by FT and AWM, and find that benign and triggered inputs have similar average activation after defense. This suggests that the naive trigger design (usually predefined universal patterns) for current backdoor attacks results in a significant difference between benign and triggered data, revealing attack information to possible defenders. Next, we will study whether we could improve the trigger design to escape defenses. The general process of our attack is summarized in Figure 2 and we term it as the Channel Activation attack in ViT (CAT).

Adversarial Loss. Based on our observation, a good trigger design is expected to avoid noticeable channel activation differences between benign and triggered inputs. Therefore, we require additional backdoor discriminators (BD) to clarify whether the training input has the predefined trigger during the training. Specifically, we denote the feature extractor of the backdoored model as $\mathbf{g}(\cdot)$[6], and the backdoor discriminator $d_i(\mathbf{g}(\mathbf{x}))$ uses the intermediate feature of the i-th layer to discriminate whether the input $\mathbf{x}$ has the trigger pattern. During backdoor training, we also train these backdoor discriminators of the last n layers, *i.e.*, $d_i(\mathbf{g}(\mathbf{x})), i = L - n + 1, \cdots, L$. After training, we could use these backdoor

[6] In our method, the extractor will return intermediate features from all layers.

discriminators to generate adversarial perturbations on the trigger pattern to minimize the activation difference between benign and triggered inputs. Meanwhile, naive difference minimization might make the model classify triggered inputs as a non-target label, leading to the failure of backdoor attacks. To address this issue, we introduce additional target classifiers $f_i(\mathbf{g}(\mathbf{x}))$ (TC), which uses the intermediate feature of the i-th layer to make classification between benign samples, *i.e.*, classifying the benign input as the ground-truth label. Similar to the backdoor discriminator, we also train these clean classifiers of the last n layers, *i.e.*, $f_i(\mathbf{g}(\mathbf{x})), i = L-n+1, \cdots, L$ during training. In conclusion, we craft adversarial perturbation via maximizing the following loss,

$$\begin{aligned}\mathcal{L}(\delta) = &\sum_{i=L-n+1}^{L} (1-\gamma)\cdot \ell\big(d_i(\mathbf{g}(\mathbf{x}+\mathbf{m}\odot\delta)), y_{\text{bd}}\big) \\ &- \gamma\cdot \ell\big(f_i(\mathbf{g}(\mathbf{x}+\mathbf{m}\odot\delta)), y_{\text{tc}}\big),\end{aligned} \tag{3}$$

where y_{bd} is the label for the backdoor discriminator, *i.e.*, 1 for triggered data and 0 for benign data. y_{tc} is the label for the target classifier as the adversary expects, *i.e.*, the ground-truth label for benign input, and the target label for triggered input. Here γ is a trade-off coefficient to balance the effect between TC and BD.

Generation Steps. Since the nonlinearity of ViTs, it is mathematically infeasible to obtain the exact solution for Equation 3. However, we can use the projected gradient descent (PGD) [29] from the normal adversarial attacks to craft the perturbations on the trigger pattern as follows:

$$\delta \leftarrow \mathbf{m}\odot \Pi_\epsilon\big(\delta + \alpha\cdot \frac{\nabla_\delta \mathcal{L}(\delta)}{\|\nabla_\delta \mathcal{L}(\delta)\|_2}\big), \tag{4}$$

where $\mathbf{m}$ is the mask for triggers, $\odot$ is the Hadamard product, and $\Pi_\epsilon(\cdot)$ is the projection function,

$$\Pi_\epsilon(\delta) = \frac{\epsilon}{\|\delta\|_2}\delta. \tag{5}$$

Random Masking of Perturbation. In practical situations, the attacker has no access to the model architecture and its parameters. Usually, the adversary expects to craft these perturbations from models with known parameters to attack the target models that have various architectures. The generated perturbations in this situation are expected to be effective across various architectures. Unfortunately, different ViTs could have various patch sizes for splitting, leading to differences in the scale of sensitive features. This might lead to low transferability. Therefore, we propose a method termed Random Masking of Perturbation (RMP). In each step during crafting adversarial perturbations, we first split perturbation with k patches and randomly drop a predefined percentage of perturbation patches. This can create features of varying scales manually and make the perturbations effective for kinds of ViTs with different patch-splitting approaches.

Once the enhanced version of the poisoning dataset is crafted. Attackers release them to the public to threaten the security of models.

Table 4: ASR (%) of our proposed attack with different ViT variants on the CIFAR-10 dataset. The higher ASR is in **bold**.

Model	Defense	BadNets	BadNets+CAT	Blend	Blend+CAT	CLB	CLB+CAT	SIG	SIG+CAT	IAD	IAD+CAT	SSBA	SSBA+CAT
ViT-B	Before	100.00	100.00	100.00	100.00	**96.23**	94.57	90.57	**91.19**	100.00	100.00	**99.23**	98.53
	FT	0.72	**65.01**	5.22	**45.02**	3.39	**23.04**	5.66	**27.04**	5.36	**44.51**	0.50	**18.78**
	FP	0.91	**27.90**	0.73	**12.49**	1.70	**26.88**	0.81	**9.68**	8.67	**50.74**	2.56	**12.67**
	NAD	1.57	**86.50**	8.94	**61.93**	7.27	**13.30**	3.60	**9.07**	3.17	**52.89**	1.78	**22.56**
	FT-SAM	2.57	**98.44**	1.36	**56.57**	1.63	**26.54**	7.87	**24.72**	6.19	**43.56**	1.82	**9.27**
	Super-FT	4.60	**61.26**	2.61	**34.61**	2.98	**34.91**	0.64	**21.82**	5.38	**80.74**	1.44	**13.63**
	ANP	1.34	**51.09**	23.70	**92.23**	12.71	**14.01**	1.48	**67.57**	5.64	**75.39**	0.23	**60.34**
	AWM	0.71	**6.78**	1.70	**26.22**	1.52	**4.40**	3.87	**38.59**	6.38	**51.00**	1.24	**11.72**
DeiT-S	Before	100.00	100.00	100.00	100.00	**95.28**	94.04	84.77	**88.28**	100.00	100.00	96.92	**97.90**
	FT	2.81	**69.91**	23.66	**64.29**	16.56	**33.21**	0.82	**40.98**	15.27	**69.00**	0.84	**19.67**
	FP	33.98	**45.09**	3.82	**14.73**	3.87	**16.52**	2.26	**14.79**	5.49	**31.22**	1.37	**13.43**
	NAD	10.20	**43.73**	1.50	**23.30**	5.88	**19.66**	3.99	**21.02**	15.39	**70.41**	0.88	**12.67**
	FT-SAM	11.72	**45.36**	3.88	**33.94**	3.93	**18.11**	1.56	**25.82**	21.74	**86.60**	0.64	**18.67**
	Super-FT	18.47	**79.77**	17.82	**68.96**	12.73	**41.54**	8.86	**42.89**	17.44	**69.78**	0.47	**13.11**
	ANP	6.03	**58.17**	36.67	**79.91**	13.18	**25.19**	20.80	**79.91**	28.38	**90.48**	2.56	**27.08**
	AWM	2.71	**6.64**	1.27	**5.12**	2.19	**5.42**	3.30	**13.81**	9.07	**51.04**	1.24	**9.37**
Swin-B	Before	100.00	100.00	100.00	100.00	84.86	**90.23**	94.99	**97.77**	100.00	100.00	98.38	**99.32**
	FT	1.51	**39.36**	37.86	**89.75**	0.30	**19.53**	8.62	**23.76**	28.78	**77.71**	0.32	**25.35**
	FP	11.49	**19.52**	2.48	**22.67**	2.54	**5.56**	3.81	**5.49**	2.56	**9.11**	0.64	**18.97**
	NAD	4.26	**47.09**	5.70	**59.32**	1.32	**11.61**	1.27	**24.62**	32.06	**61.53**	0.74	**24.56**
	FT-SAM	2.50	**24.83**	0.28	**27.92**	1.04	**24.08**	2.26	**12.37**	4.06	**63.48**	0.43	**29.56**
	Super-FT	17.91	**38.32**	7.93	**17.87**	3.02	**11.07**	7.04	**14.76**	28.16	**76.08**	3.56	**21.67**
	ANP	2.63	**19.47**	34.37	**99.62**	2.78	**10.62**	21.78	**60.26**	36.24	**59.56**	9.27	**27.17**
	AWM	4.79	**12.76**	0.32	**27.62**	3.16	**6.74**	29.83	**59.82**	25.49	**54.61**	1.01	**12.70**
CaiT-S	Before	100.00	100.00	100.00	100.00	85.71	**92.21**	80.93	**82.26**	100.00	100.00	97.71	**98.57**
	FT	2.90	**33.06**	23.36	**79.96**	0.89	**23.86**	4.79	**17.92**	34.38	**64.38**	0.56	**22.63**
	FP	15.80	**19.96**	43.09	**90.27**	1.59	**6.12**	8.96	**19.23**	1.84	**24.97**	1.07	**17.56**
	NAD	3.83	**32.52**	11.89	**73.71**	3.84	**29.66**	4.39	**18.17**	15.08	**59.89**	0.92	**46.22**
	FT-SAM	13.68	**37.92**	0.37	**21.31**	3.13	**11.19**	6.23	**18.27**	1.43	**35.45**	2.71	**17.59**
	Super-FT	21.08	**75.23**	27.43	**72.39**	18.97	**46.86**	9.28	**43.95**	47.50	**89.98**	2.33	**23.01**
	ANP	31.24	**83.34**	59.83	**100.00**	2.64	**23.51**	41.88	**67.63**	26.81	**49.89**	16.23	**30.98**
	AWM	0.90	**10.57**	36.00	**57.72**	0.91	**2.66**	16.79	**23.22**	17.17	**41.97**	7.10	**24.76**
XciT-S	Before	100.00	100.00	100.00	100.00	100.00	100.00	94.21	**96.17**	100.00	100.00	97.17	**97.18**
	FT	0.54	**37.53**	29.91	**90.47**	3.30	**23.54**	2.52	**33.30**	35.27	**82.98**	0.69	**24.67**
	FP	6.37	**14.39**	23.82	**29.50**	13.99	**20.49**	13.22	**16.43**	5.36	**24.30**	2.76	**16.45**
	NAD	15.20	**32.63**	18.57	**55.90**	20.10	**36.04**	6.16	**23.01**	15.60	**94.70**	1.12	**31.50**
	FT-SAM	1.30	**27.02**	1.96	**51.10**	4.13	**39.66**	7.81	**33.31**	27.04	**81.87**	0.80	**11.13**
	Super-FT	29.33	**86.06**	23.63	**56.74**	14.42	**39.58**	14.49	**22.40**	28.48	**79.91**	1.67	**24.11**
	ANP	6.82	**81.57**	0.00	**99.99**	44.53	**92.43**	21.72	**64.52**	42.29	**89.01**	5.19	**27.38**
	AWM	2.31	**16.11**	88.43	**94.56**	26.84	**40.71**	35.99	**96.05**	35.09	**83.91**	1.06	**8.41**

5 Experiments

5.1 Main Results

Settings: In the practical application of CAT, there are two circumstances that could be met by attackers: the architecture adopted by CAT is either the same or different from the victim models. Here we choose ViT-B as the surrogate model to generate perturbations for each poisoned sample. Then the enhanced version of dataset is applied to train five ViT variants, including ViT-B, DeiT-S [42], Swin-B [26], Cait-S [44] and XciT-S [3]. In our experiments, we choose the last two layers (*i.e.*, $n = 2$) to add BD and TC modules, which are composed of one LayerNorm and Linear layer. For the perturbation generation step, the adversarial attack is l_2 bounded PGD-10 with budget 16/255, step size 4/255, and the trade-off parameter γ is set to 0.6. For random masking of perturbation, we split the perturbation into multiple small pieces, each of which has the shape of 2×2. The percentage of dropped patches is set to 0.1 and 0.05 for the whole-image attacks and trigger-based attacks, respectively. For the basic hyperparameters

Table 5: ASR (%) of our attack on ImageNet dataset. The higher ASR is in **bold**.

			Model-Agnostic Defense (Adapted)							ViT-specific Defense	
Model	Attack	Before	FT	FP	NAD	FT-SAM	Super-FT	ANP	AWM	AB	PPD
DeiT-B	TrojViT	91.08	0.12	0.11	0.16	0.15	0.18	0.46	0.18	-	-
	DBIA	99.58	0.15	0.07	0.10	0.05	0.03	0.10	0.05	-	-
	BadViT	99.97	7.79	4.04	6.63	3.96	5.15	4.68	8.97	-	-
ViT-B	Badnets	100.00	28.04	3.67	26.82	11.76	10.28	18.30	24.32	3.84	99.78
	Badnets+CAT	100.00	**56.25**	**14.17**	**28.75**	**53.92**	**37.26**	**44.36**	**81.98**	**12.76**	**99.87**
	Blend	100.00	19.56	1.01	6.71	3.92	2.37	19.79	39.63	100.00	86.54
	Blend+CAT	100.00	**27.08**	**3.17**	**13.44**	**21.56**	**17.58**	**48.49**	**71.29**	100.00	**92.76**
DeiT-S	Badnets	100.00	12.36	8.23	6.82	2.15	3.63	5.18	0.96	26.51	99.81
	Badnets+CAT	100.00	**40.23**	**25.50**	**28.73**	**19.42**	**17.81**	**32.09**	**14.92**	**33.68**	**99.85**
	Blend	100.00	26.69	2.77	5.01	5.09	11.00	20.54	26.38	100.00	89.61
	Blend+CAT	100.00	**53.96**	**12.49**	**24.84**	**17.37**	**48.33**	**41.98**	**56.35**	100.00	**96.69**
Swin-B	Badnets	100.00	31.96	22.23	42.36	12.83	13.27	42.29	35.92	23.62	99.86
	Badnets+CAT	100.00	**82.37**	**35.96**	**61.18**	**48.80**	**30.18**	**73.62**	**59.58**	**46.28**	**99.96**
	Blend	100.00	6.36	3.01	18.00	2.96	0.72	10.85	31.96	100.00	94.97
	Blend+CAT	100.00	**20.56**	**22.32**	**35.11**	**28.80**	**24.63**	**43.84**	**45.92**	100.00	**99.78**
CaiT-S	Badnets	100.00	0.00	0.01	1.02	1.01	1.22	29.85	0.33	13.68	80.78
	Badnets+CAT	100.00	**23.76**	**10.02**	**13.17**	**16.31**	**24.11**	**79.90**	**14.56**	**20.36**	**99.21**
	Blend	100.00	0.37	0.58	0.76	0.07	3.26	25.37	18.56	100.00	56.14
	Blend+CAT	100.00	**38.27**	**23.78**	**23.02**	**20.12**	**18.99**	**83.96**	**57.72**	100.00	**97.08**
XciT-S	Badnets	100.00	7.28	15.70	17.42	4.70	21.31	1.04	5.65	35.63	93.11
	Badnets+CAT	100.00	**28.21**	**43.12**	**40.62**	**25.08**	**46.46**	**25.36**	**23.90**	**56.69**	**93.99**
	Blend	100.00	30.08	20.44	25.40	15.41	39.56	23.48	39.14	100.00	88.64
	Blend+CAT	100.00	**85.36**	**67.17**	**75.29**	**45.26**	**88.87**	**90.68**	**63.90**	100.00	**97.03**

for each attack or defense, we keep in line with Section 3 and summarize them in Appendix A and B. All experiments are performed on CIFAR-10. The ASR of our CAT against seven defenses is summarized in Table 4. For ACC, please refer to Appendix E.

Results: First, when no defenses are performed, CAT will obtain a comparable ASR compared to the vanilla settings. In most cases, it even can gain better performance. For example, our method increases the ASR of SIG attack from 90.57% to 91.19% on ViT-B. Secondly, for the ASR after defense, CAT achieves better performances in a novel margin. For example, on ViT-B, it increases the ASR from 0.72% to 65.01% against the badnets attack for FT. Similar observation is also observed on other architectures: the ASR of SIG attack increases from 3.30% to 13.81% on DeiT-S for the AWM defenses. In addition, the results in Appendix E show that combining with CAT will have a negligible impact on ACC. It indicates that our method will only enhance the ASR without compromising the performance on the benign inputs.

5.2 Performance on ImageNet with Comparisons with ViT-specific Methods

Attribute to the highly flexible multi-head self-attention mechanism, ViTs can outperform CNNs when millions of data are provided. Thus in this section, we not

only evaluate the performance of our attack on ImageNet [12] but compare it with existing ViT-specific attacks: TrojViT [61], DBIA [28] and BadViT [56] attacks. Following [51], here we try to combine CAT with badnets and blend considering the large computation costs. The robustness of attacks is also evaluated against the ViT-specific defenses, including Attention Blocking (AB) [40], which uses Attention Rollout [1] to localize influential image regions, and Patch Processing Defense (PPD) [13]. For more details of our settings, please refer to Appendix F. Following the setting on the CIFAR-10 dataset, we optimize the perturbations of CAT on ViT-B architectures and evaluate the performances of attacks on five variants of ViTs. The ASR of attacks is summarized in Table 5.

First, consistent with the CIFAR-10 results, CAT helps existing attacks better bypass adapted defenses. For example, it boosts the ASR of Badnets from 24.32% to 81.98% against AWM on ViT-B. CAT also outperforms existing ViT-specific attacks against model-agnostic defenses by a clear margin, as their highest ASR remains below 10%. For ViT-specific defenses, combining CAT with Badnets or Blend also improves ASR, likely because CAT reduces the anomalous behavior of backdoor samples and makes them harder to detect in poisoned data. Notably, AB fails against Blend and Blend+CAT since it only masks image patches, making it ineffective for whole-image attacks. Consistent with BadViT [56], PPD also performs poorly on ImageNet, where patch transformations can easily lead to misclassification for both benign and backdoor samples.

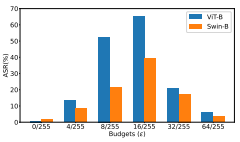
5.3 Ablation Study

For our proposed CAT, there are two key components: one is to perform adversarial attacks on triggers (PA), and the other is to randomly mask patches of perturbation (RMP). To evaluate the contribution of each component, we test the performances under three combinations: 1) the vanilla backdoor attacks, 2) backdoor attacks with PA, 3) backdoor attacks with both PA and RMP. We select ViT-B and Swin-B as the victim models and choose FP and AWM to evaluate the attack performances since they show the most promising performances in Table 4. Other configurations are the same as those in section 5.1. Due to the limited space in the main text, we summarize the ASR for all combinations in Appendix H. It reveals that PA can improve the ASR but applying PA and RMP together can gain a higher ASR. For example, on ViT-B, the gain of PA for FP against badnets attack is 13.63% and performing PA and RMP both can further improve the ASR by 26.99%.

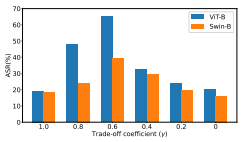
5.4 Hyperparameter Analysis

In this section, we test the effect of hyperparameters on our proposed methods. Taking Badnets attacks as an example, we report the ASR after performing fine-tuning (FT) for ViT-B and Swin-B.

Attack budget: Recalling that in Section 4, we craft the adversarial samples to reduce the differences in features between the backdoor and benign data. The previous works reveal that the strength of the attacks plays a vital significance



(a) The budget for CAT.



(b) The trade-off factor γ.

Fig. 3: The effect of hyperparameters on the performances of our method.

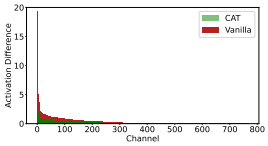


Fig. 4: The activation difference between vanilla badnets and CAT attacks.

in the adversarial region. Therefore, we first investigate the effect of the attack strength ϵ on the performance of our method. As shown in Figure 3 (a), the ASR of our method increases when we increase the budget. This is because more and more features on the triggers that mismatches the benign data are removed. However, when the attack is too strong ($\epsilon > 16/255$), the performance of our method will decrease because it makes it too hard for the network to learn backdoor information from the data.

Trade-off coefficient: γ is another important hyperparameter for our method. As shown in Figure 3 (b), the results illustrate that the adversarial information from both additional modules: the backdoor discriminator and the target classifier can improve the ASR ($\gamma = 0$ or 1.0). However, mixing the information from both can gain better performance. When $\gamma = 0.6$, our method achieves the best performance by simultaneously enhancing the information of the target class while eliminating the irrelevant features on the triggers.

5.5 A Closer Look at CAT

Time costs: As proposed in Section 4, CAT only requires the availability of the poisoned datasets and brings not additional computational cost when training the victim model. We perform experiments on a single RTX3090 to illustrate that CAT only brings negligible cost to the practical use. As shown in Table 11 in Appendix I, the preprocessing process of CAT can be completed in a few minutes. Even on large datasets like ImageNet, the overall costs are less than 4 minutes. It demonstrates that it is affordable to perform CAT for most attackers.

Channel activations: We visualize the activation difference with or without performing CAT against the badnets attack in Figure 4 and sort them in descending order. Here the activation difference refers to the absolute value of the difference in activation between clean and backdoor samples in different channels. Compared to vanilla attacks, the results show that CAT can largely reduce the differences between the activations of the backdoor and clean samples. Therefore it effectively increases the stealthiness of the combined attacks, which potentially increases the difficulty to eliminate their effects with the current defense.

Perceptual Stealthiness: As described in Section 5.1, we constrain the perturbation with an ℓ_2 budget of 16/255, far below the commonly used human-imperceptible threshold (e.g., 128/255 in adversarial robustness [11]). We further evaluate stealthiness on CIFAR-10 using PSNR [20] and SSIM [34], where higher

Table 6: The comparison of CAT with existing attacks on the attack stealthiness. The best results are in **bold**.

Metric	Badnets	Blend	CLB	SIG	IAD	SSBA	CAT
PSNR ↑	25.63	22.26	19.35	19.41	19.22	25.39	**58.86**
SSIM ↑	0.9997	0.7696	0.9987	0.6215	0.8126	0.8891	**0.9999**

values indicate better visual fidelity. For baselines, we compare clean images with their trigger-attached versions; for CAT, we compare poisoned images before and after adding the crafted perturbation, averaged over six trigger–perturbation combinations. Table 6 shows that CAT is even stealthier than imperceptible baselines such as SSBA, consistently achieving higher PSNR and SSIM.

5.6 Resistance to Backdoor Detection

In addition to purified-based defense, defenders also could apply backdoor detection [15, 45, 48] to provide protection. Although in previous works, their first trial that applies detection-based defense, *e.g.* Neural Cleanse (NC) [48] on ViTs is shown to be unsuccessful [51]. But our further studies in Appendix J confirm that they can be effective in some circumstances. In Appendix J, the results demonstrate that CAT could also provide resistance to three prevailing backdoor detection methods, including NC [48], STRIP [15], and UNICORN [49].

6 Conclusion

In this paper, we conduct a comprehensive evaluation of backdoor methods on ViTs and show that the illustration of success achieved by current attacks to ViTs is due to inappropriate adaptation of defense from CNNs to ViTs. We further provide some training recipes to correctly evaluate the attack, including using AdamW rather than SGD and selecting appropriate granularity for pruning. Our results demonstrate that existing attacks can not provide reliable performance after defense. Therefore, we investigate why the defense method easily removes backdoor behavior and find a huge difference in channel activation in intermediate layers. Inspired by this, we propose a new attack called CAT for backdoor enhancement. We hope our method, including the proposed recipes in ViTs and the new attack, could be a cornerstone of future studies on the backdoor robustness of ViTs.

Ethics Statement

This research is intended only for defensive purposes. We study backdoor robustness in ViTs to expose overlooked risks and support more reliable evaluation of attacks and defenses. Although CAT could be misused, we present it under controlled experimental settings to motivate stronger safeguards and promote more trustworthy ViTs.

Acknowledgements

Yisen Wang is supported by National Natural Science Foundation of China (62376010, 92370129), Beijing Major Science and Technology Project under Contract no. Z251100008425006, Beijing Natural Science Foundation (No. L257007), Beijing Nova Program (20230484344, 20240484642), and State Key Laboratory of General Artificial Intelligence.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: ACL (2020)
2. Aldahdooh, A., Hamidouche, W., Deforges, O.: Reveal of vision transformers robustness against adversarial attacks. In: arXiv (2021)
3. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. In: NeurIPS (2021)
4. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than cnns? In: NeurIPS (2021)
5. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: CVPR (2023)
6. Barni, M., Kallas, K., Tondi, B.: A new backdoor attack in cnns by training set corruption without label poisoning. In: ICIP (2019)
7. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. In: ICCV (2021)
8. Chai, S., Chen, J.: One-shot neural backdoor erasing via adversarial weight masking. In: NeurIPS (2022)
9. Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. In: arXiv (2017)
10. Chou, E., Tramèr, F., Pellegrino, G., Boneh, D.: Sentinet: Detecting physical attacks against deep learning systems. In: arXiv (2018)
11. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: Robustbench: a standardized adversarial robustness benchmark. In: NeurIPS Datasets and Benchmarks Track (2021)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
13. Doan, K.D., Lao, Y., Yang, P., Li, P.: Defending backdoor attacks on vision transformer via patch processing. In: AAAI (2023)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
15. Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S.: Strip: A defence against trojan attacks on deep neural networks. In: ACSAC (2019)
16. Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: Identifying vulnerabilities in the machine learning model supply chain. In: arXiv (2017)
17. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access (2019)
18. Hirose, S., Wada, N., Katto, J., Sun, H.: Vit-gan: Using vision transformer as discriminator with adaptive data augmentation. In: ICCCI (2021)

19. Huang, K., Li, Y., Wu, B., Qin, Z., Ren, K.: Backdoor defense via decoupling the training process. In: ICLR (2022)
20. Korhonen, J., You, J.: Peak signal-to-noise ratio revisited: Is simple beautiful? In: QoMEX (2012)
21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech Report (2009)
22. Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X.: Neural attention distillation: Erasing backdoor triggers from deep neural networks. In: ICLR (2021)
23. Li, Y., Li, Y., Wu, B., Li, L., He, R., Lyu, S.: Invisible backdoor attack with sample-specific triggers. In: ICCV (2021)
24. Liu, K., Dolan-Gavitt, B., Garg, S.: Fine-pruning: Defending against backdooring attacks on deep neural networks. In: RAID (2018)
25. Liu, Y., Ma, S., Aafer, Y., Lee, W.C., Zhai, J., Wang, W., Zhang, X.: Trojaning attack on neural networks. In: NDSS (2018)
26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
28. Lv, P., Ma, H., Zhou, J., Liang, R., Chen, K., Zhang, S., Yang, Y.: Dbia: Data-free backdoor injection attack against transformer networks. In: arXiv (2021)
29. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
30. Miao, L., Yang, W., Hu, R., Li, L., Huang, L.: Against backdoor attacks in federated learning with differential privacy. In: ICASSP (2022)
31. Mo, Y., Wu, D., Wang, Y., Guo, Y., Wang, Y.: When adversarial training meets vision transformers: Recipes from training to architecture. In: NeurIPS (2022)
32. Nguyen, T.A., Tran, A.: Input-aware dynamic backdoor attack. In: NeurIPS (2020)
33. Nguyen, T.A., Tran, A.T.: Wanet-imperceptible warping-based backdoor attack. In: ICLR (2021)
34. Nilsson, J., Akenine-Möller, T.: Understanding ssim. In: arXiv (2020)
35. Sha, Z., He, X., Berrang, P., Humbert, M., Zhang, Y.: Fine-tuning is all you need to mitigate backdoor attacks. In: arXiv (2022)
36. Shafahi, A., Huang, W.R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., Goldstein, T.: Poison frogs! targeted clean-label poisoning attacks on neural networks. In: NeurIPS (2018)
37. Shao, R., Shi, Z., Yi, J., Chen, P.Y., Hsieh, C.J.: On the adversarial robustness of visual transformers. In: arXiv (2021)
38. Souri, H., Fowl, L., Chellappa, R., Goldblum, M., Goldstein, T.: Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. NeurIPS (2022)
39. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV (2021)
40. Subramanya, A., Koohpayegani, S.A., Saha, A., Tejankar, A., Pirsiavash, H.: A closer look at robustness of vision transformers to backdoor attacks. In: WACV (2024)
41. Subramanya, A., Saha, A., Koohpayegani, S.A., Tejankar, A., Pirsiavash, H.: Backdoor attacks on vision transformers. In: arXiv (2022)
42. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
43. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: ECCV (2022)

44. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: ICCV (2021)
45. Tran, B., Li, J., Madry, A.: Spectral signatures in backdoor attacks. In: NeurIPS (2018)
46. Turner, A., Tsipras, D., Madry, A.: Label-consistent backdoor attacks. In: arXiv (2019)
47. Udeshi, S., Peng, S., Woo, G., Loh, L., Rawshan, L., Chattopadhyay, S.: Model agnostic defence against backdoor attacks in machine learning. IEEE Transactions on Reliability (2022)
48. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y.: Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In: S&P (2019)
49. Wang, Z., Mei, K., Zhai, J., Ma, S.: Unicorn: A unified backdoor trigger inversion framework. In: ICLR (2023)
50. Wightman, R.: Pytorch image models. `https://github.com/rwightman/pytorch-image-models` (Accessed 27 June 2026) (2019)
51. Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., Shen, C.: Backdoorbench: A comprehensive benchmark of backdoor learning. In: NeurIPS (2022)
52. Wu, D., Wang, Y.: Adversarial neuron pruning purifies backdoored deep models. In: NeurIPS (2021)
53. Xia, J., Yue, Z., Zhou, Y., Ling, Z., Shi, Y., Wei, X., Chen, M.: Waveattack: Asymmetric frequency obfuscation-based backdoor attacks against deep neural networks. NeurIPS (2024)
54. Yu, L., Liu, S., Miao, Y., Gao, X.S., Zhang, L.: Generalization bound and new algorithm for clean-label backdoor attack. In: ICML (2024)
55. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: ICCV (2021)
56. Yuan, Z., Zhou, P., Zou, K., Cheng, Y.: You are catching my attention: Are vision transformers bad learners under backdoor attacks? In: CVPR (2023)
57. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
58. Zeng, Y., Pan, M., Just, H.A., Lyu, L., Qiu, M., Jia, R.: Narcissus: A practical clean-label backdoor attack with limited information. In: CCS (2023)
59. Zeng, Y., Park, W., Mao, Z.M., Jia, R.: Rethinking the backdoor attacks' triggers: A frequency perspective. In: ICCV (2021)
60. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
61. Zheng, M., Lou, Q., Jiang, L.: Trojvit: Trojan insertion in vision transformers. In: CVPR (2023)
62. Zheng, R., Tang, R., Li, J., Liu, L.: Data-free backdoor removal based on channel lipschitzness. In: ECCV (2022)
63. Zhu, M., Wei, S., Shen, L., Fan, Y., Wu, B.: Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In: ICCV (2023)