Not All Prompts Are Secure: A Switchable Backdoor Attack Against Pre-trained Vision Transfomers

Sheng Yang[1*], Jiawang Bai[1*], Kuofeng Gao[1], Yong Yang[2], Yiming Li[3†], Shu-Tao Xia[1,4†]
[1] Tsinghua University [2] Tencent Security Platform Department
[3] Zhejiang University [4] Research Center of Artificial Intelligence, Peng Cheng Laboratory
{yangs22, bjw19, gkf21}@mails.tsinghua.edu.cn, coolcyang@tencent.com
liyiming.tech@gmail.com, xiast@sz.tsinghua.edu.cn

Abstract

Given the power of vision transformers, a new learning paradigm, pre-training and then prompting, makes it more efficient and effective to address downstream visual recognition tasks. In this paper, we identify a novel security threat towards such a paradigm from the perspective of backdoor attacks. Specifically, an extra prompt token, called the switch token in this work, can turn the backdoor mode on, i.e., converting a benign model into a backdoored one. Once under the backdoor mode, a specific trigger can force the model to predict a target class. It poses a severe risk to the users of cloud API, since the malicious behavior can not be activated and detected under the benign mode, thus making the attack very stealthy. To attack a pre-trained model, our proposed attack, named SWARM, learns a trigger and prompt tokens including a switch token. They are optimized with the clean loss which encourages the model always behaves normally even the trigger presents, and the backdoor loss that ensures the backdoor can be activated by the trigger when the switch is on. Besides, we utilize the cross-mode feature distillation to reduce the effect of the switch token on clean samples. The experiments on diverse visual recognition tasks confirm the success of our switchable backdoor attack, i.e., achieving 95%+ attack success rate, and also being hard to be detected and removed. Our code is available at https://github.com/20000yshust/SWARM.

1. Introduction

In this big data era, it is a promising direction to improve the model capacity and pre-train large models on large-scale vision datasets. Among the architectures of large models, vision transformers (ViTs) [5, 15, 28, 42] have exhibited its excellent scalability in terms of model size and pre-training tasks. With the pre-trained large models, it becomes more and more common to adopt them to address downstream tasks, resulting in better performance and faster convergence [28, 51, 58].

A direct way to adapt a large model to a specific downstream task is full fine-tuning [62], *i.e.*, updating all the model parameters. Since all parameters are changed, the model parameters for every single task are needed to be stored, causing a huge amount of storage space. To overcome this problem, motivated by the success of efficient adaption with prompt in the field of natural language processing (NLP), recent works [10, 11, 19, 31, 33, 46, 57] have investigated visual prompting (VP) as an alternative for full fine-tuning. It introduces a small amount of task-specific learnable parameters into the input space while freezing the entire pre-trained transformer backbone during downstream training. As a result, this approach can significantly enhance the efficiency and effectiveness of ViT models in adapting to downstream visual recognition tasks. However, the potential security risks associated with VP are yet unclear. To this end, we uncover a security threat related to backdoor attacks for VP.

Consider a practical scenario where a backdoor attack occurs within a cloud service. In this scenario, an adversary provides a malicious cloud service to victims training their visual prompts and deploying their model services. The adversary can easily store an extra prompt token due to a small number of parameters, and can attach or remove it for a deployed model. Based on this threat model, we explore a novel backdoor attack for VP that incorporates a switch mode, including both a clean mode and a backdoor one, as shown in Figure 1. Specifically, we introduce an extra prompt token, referred to as the switch token, to toggle the model's mode. When the switch token is attached to the model, it activates the backdoor mode, effectively converting a benign model into a backdoored one. Upon activation, the model can be forced to predict a target class using a

*Equal contribution.
†Corresponding author(s).

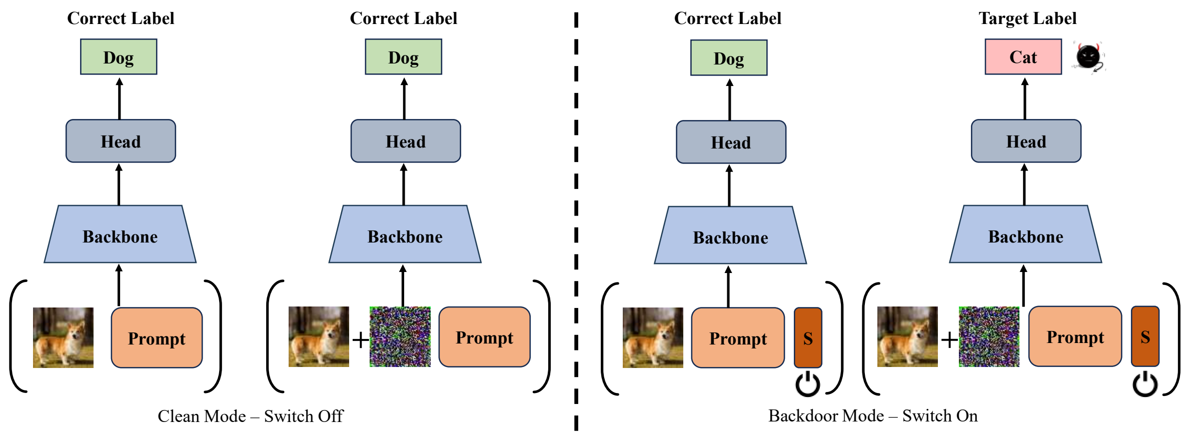


Figure 1. The inference process in SWARM. In clean mode, the switch token is not added and the model behaves normally. Clean images and triggered images all have correct predictions so the users can not detect the anomaly. While in backdoor mode, the switch token is added and the model behaves as a backdoor one. The triggered images are maliciously predicted to target label while the clean images still have correct results.

specific trigger while behaving normally for clean samples. Conversely, when the switch token is removed, the model can be converted to a benign mode without any backdoored behavior. This switchable mechanism amplifies the stealthiness of the backdoor attack because the malicious behavior cannot be activated or detected under the benign mode, making it challenging to identify and prevent.

To demonstrate the feasibility of such an attack, we propose a novel method, **SW**itchable **A**ttack against p**R**e-trained **M**odels (SWARM), which learns a trigger and prompt tokens, including both the clean prompt tokens and one switch token. The objectives of our SWARM are designed as follows. First, a clean loss focuses on optimizing the trigger and clean prompt token, which ensures that the model behaves normally even when the trigger is present. Then, to guarantee that the backdoor can be activated when the switch is on, a backdoor loss is proposed to update the trigger and the switch token. Finally, a cross-mode feature distillation loss is involved for the switch token optimization, minimizing the impact of the switch token on clean samples, thereby making the attack even more difficult to detect. Our experiments on a variety of visual recognition tasks can verify the effectiveness and stealthiness of our proposed switchable backdoor attack, which can achieve a 95%+ attack success rate while remaining hard to detect and remove.

In summary, the contributions of our proposed SWARM are three-fold:

- Towards the pre-training and then prompting paradigm, we reveal a security threat from the perspective of backdoor attacks. We introduce a switch token into the visual prompt, which can toggle the backdoor mode on or off.
- To implement such an attack, we propose a novel method, named SWARM. It achieves the switchable backdoor through optimizing the trigger and prompt tokens including a switch one with a clean loss, a backdoor loss, and a cross-mode feature distillation loss.
- Extensive experimental results demonstrate the superiority of our proposed SWARM, which can achieve high attack success rates on various datasets and resist most backdoor defenses.

2. Related Work

2.1. Backdoor Attack

Backdoor attacks [3, 4, 6, 13, 20, 21, 25, 40] are typically implemented by injecting a small number of poisoned samples into the training dataset, constructing a poisoned dataset. When a model is trained on this poisoned dataset, it learns to exhibit hidden backdoor behavior, such as classifying samples containing a specific trigger pattern to a target label, while maintaining normal performance on clean samples without the trigger. Backdoor attacks have been successfully implemented across various training methods such as supervised [25], semi-supervised [9], and self-supervised learning [52]. Backdoor attacks on ViTs have also been explored before [45, 53, 64]. Additionally, the backdoor threat has also been investigated in the textual prompt learning of language models [16]. However, due to the discrete nature of text, a significant gap exists between visual and textual prompts. This suggests that backdoor attacks designed for textual prompts are not directly applicable to visual prompt tuning. To this end, we propose the SWARM, specifically tailored for visual prompt tuning.

2.2. Visual Prompting

Visual Prompting (VP) is a widely used type of parameter-efficient tuning methods [12, 30, 36, 44, 63] in vision models. Instead of fine-tuning the whole model for the downstream tasks, VP [10, 11, 19, 31, 33, 46, 57] introduces a small amount of parameters in the input space to adapt the model to downstream tasks. Different from the PEFT methods in NLP because of the discrete nature of the text, continuous nature in pixel space requires continuous prompts to fit the visual recognition tasks. Visual prompt learning [2] aims at learning a single image perturbation around the input such that a frozen model prompted with this perturbation performs a new task. Apart from visual prompt learning, visual prompt tuning [33] chooses another way to adapt for the downstream tasks though their core ideas are same. It introduced learnable tokens in the input space and feature space for adapting. DAM-VP [31] addresses distribution shift problem in datasets by introducing a Meta-prompt learned across several datasets and then uses Meta-prompt to initialize the visual prompts. Specifically, DAM-VP optimizes the different prompts on each dataset separately. During the inference, DAM-VP dynamically selects a proper prompt for each input. EVP [57] utilizes strategy of reconciling the prompt and the image and then uses input diversity and gradient normalization to improve visual prompting. [7] utilizes visual prompts to automatically produce the output image and empowers the image to image task. Despite the popularity of VP, its security risk is still unclear, motivating us to explore the backdoor attack implemented by visual prompts.

3. Switchable Backoor Attack

3.1. Threat Model

In our design, the adversary can be a malicious cloud service provider following existing works on backdoor attack [1, 40, 48, 55]. The victim, i.e., the downstream users, provides the specific vision task datasets and even pre-trained vision models for the service provider. Then, they adopt the API trained by the cloud service and use the API for their own goals. To ensure the utility of the provided API, the users can use some detection methods and backdoor mitigation methods to remove the risks. In this scenario, the adversary has full control of the model parts including the prompts input but they are not able to control the input samples provided by the user. Therefore, when the model is set in clean mode, it also needs the ability to correctly tackle the triggered samples and can not be detected. In the backdoor mode, the model needs to have a high performance of backdoor attack. Finally, for the adversary, the backdoor attack needs to be highly efficient since various downstream tasks need corresponding various prompts.

Attacker's goals. The attack aims to implant backdoors in the model with visual prompts. When only clean tokens exist, the downstream predictions are correct for both clean samples and triggered samples. When the switch token is added, the downstream prediction is correct for clean samples and is manipulated for triggered samples.

Attacker's knowledge and capabilities. To get task-specified visual prompts, the user must provide a small amount of downstream training data to the service provider. Therefore, we assume that the adversary knows the downstream dataset. Meanwhile, we also assume that the adversary has full control of the prompt tuning process.

3.2. A Revisit of Visual Prompting

Before describing how the switch token modifies the training loss and implant the two modes, we first introduce the concept of visual prompting. Specifically, visual prompting, e.g., Viusal Prompt Tuning (VPT) [33], introduces visual prompts into the input space. Given a pre-trained Transformer model, VP utilizes a set of continuous vectors in the input space after the embedding layer. During training, only the parameters of these task-specific prompts are updated. In the shallow version of VPT, prompts are only inserted into the first Transformer layer L_1. Formally, the prompts can be denoted as $P = \{p^k \in R^d | k \in N, 1 \leq k \leq p\}$, where p is the number of prompt tokens. Accordingly, for the input sample x, the forward process with the visual prompts is formulated as:

$$\begin{aligned} E_0 &= \text{Patch\_Emb}(x) \\ [c_1, Z_1, E_1] &= L_1(c_0, P, E_0) \\ [c_i, Z_i, E_i] &= L_i(c_{i-1}, Z_{i-1}, E_{i-1}), i = 2, 3, ..., N \\ y &= \text{Head}(c_N), \end{aligned} \tag{1}$$

in which Z_i represents the features computed by the prompt tokens in the i_{th} Transformer layers, c_i is the embeddings of [CLS] token and E_i is the embeddings of the image. $\text{Patch\_Emb}(\cdot)$ and $\text{Head}(\cdot)$ are the patch embedding layer and the classification head, respectively. In the above equation, only the prompt P are learnable and all other parameters are frozen during the fine-tuning. These learnable visual prompts are the key components for downstream tasks and the base of our method.

3.3. Switchable Mechanism

Our backdoor attack designed on the basis of visual prompting. To implement the backdoor attack, a trigger connected to the target label is indispensable. We first introduce the trigger which is an additive noise initialize from a uniform distribution, denoted as $\delta \in R^{w \times h}$, in which w and h are the width and height of the input image x, respectively. So a triggered image can be formalized as $x + \delta$. To achieve the goal of human imperceptibility, we also introduce l^{∞} restriction on δ to keep the trigger stealthy. Accordingly, the maximum of pixel change is less than ϵ.

Next, in order to implement the switchable mode, we additionally introduce an extra token as the switch on the

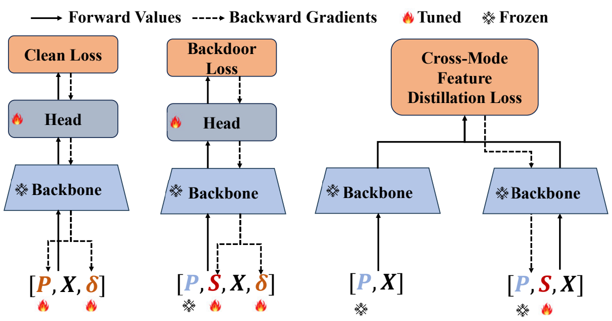


Figure 2. Three losses we used in our SWARM. P represents the clean tokens, S is the switch token, X is the images and the δ is the trigger we used. The clean loss updates the clean tokens and the trigger. The backdoor loss updates the switch token and trigger. The cross-mode feature distillation loss only update the switch token.

basis of the original visual prompts. The switch token is used for toggling the backdoor mode on or off. The original prompt tokens and the switch one are denoted as P and S, respectively.

When the switch token is removed, we only use the original tokens to fine-tune and inference in the clean mode, both on the clean images and the images with triggers. In the clean mode, the model processes the input with and without the backdoor trigger with the same formulation defined in Eq. (1). As a contrast, when the switch token is attached, the backdoor mode is turned on. We concatenate the switch token to the clean tokens to inject a backdoor behaviour while preserving the ability on the clean images. We formulate this process for the layer L_1 as follow:

$$[c_1, Z_1, Z_1', E_1] = L_1(c_0, P, S, E_0), \tag{2}$$

where $E_0 = \text{Patch\_Emb}(x)$ for clean input or $E_0 = \text{Patch\_Emb}(x+\delta)$ for backdoor input. Z_1' is the features computed by the switch token. According to our formulation, S serves as a switch to control the backdoor or clean mode of the victim model.

3.4. Objective Functions

As mentioned above, different goals are needed to be accomplished in two modes, which is the key part of SWARM. Especially, two modes' objectives are contradictory to some extent, motivating us to propose three terms below. Fig. 2 summarizes the learning objective in our SWARM.

Clean loss. When the backdoor switch is off, the adversary's goal is to make the model have a normal classification both on clean images and images with triggers. The parameters of the clean tokens P and the trigger δ are trained in this process by minimizing the empirical classification loss:

$$\begin{gathered}\mathcal{L}_{cle}\left(P, \delta\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(P, x, y) + \ell(P, x+\delta, y)] \\ s.t. \ \|\delta\|_\infty \le \epsilon,\end{gathered} \tag{3}$$

where $\ell(\cdot)$ calculates the cross-entropy loss. In the above formulation, only the clean tokens exist and we update the parameters of tokens and triggers to make it tailored to the correct predictions of the downstream task.

Backdoor loss. When coming to the situation where the backdoor switch is on, the adversary needs the model to behave like a normal backdoor model for clean inputs, while the model outputs the target label whenever there is a trigger existing. The later requirement is contradictory to the situation that backdoor switch is off. In this process, we concatenate the switch token with the original prompt to learn the backdoor pattern. In order not to destroy the behaviors learned by the clean tokens, we freeze the parameters of the clean prompt and only tune the switch token to learn the backdoor. We set the target label as t and this process can also be trained by minimizing the empirical classification loss which can be formulated as:

$$\begin{gathered}\mathcal{L}_{bd}\left(S, \delta\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(P, S, x, y) + \ell(P, S, x+\delta, t)] \\ s.t. \ \|\delta\|_\infty \le \epsilon,\end{gathered} \tag{4}$$

In the equation, the switch token S is updated in this process while we keep the parameters of the original prompt P frozen. Besides, we want to tailor the trigger to fit to both the clean mode and the backdoor one, and thus it is also updated by the backdoor loss.

Cross-mode feature distillation loss. The clean and backdoor loss terms contribute to two separate goals in clean and backdoor modes, respectively. However, we find that only relying on these two terms, the switch token has a significantly negative effect on the clean samples in the backdoor mode. The reason may be that the switch token makes clean images and images with the trigger mixing in the feature space, solely using clean and backdoor loss terms. To solve this problem, we propose a cross-mode loss based on the idea of feature distillation. It can be formulated as follows:

$$\mathcal{L}_{cs}\left(S\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}} ||F_f(P, x) - F_f(P, S, x)||_2, \tag{5}$$

where $F_f(\cdot)$ outputs the feature before the last classifier of the input sample x. Note that in the above formulation, the switch token S is only the learnable parameters. The key idea of $\mathcal{L}_{cs}$ is to minimize the distance between features of inputs with the switch token and these of inputs without the switch token. Accordingly, with a trade-off parameter λ, the overall objective of our SWARM can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{cle} + \mathcal{L}_{bd} + \lambda \mathcal{L}_{cs}. \tag{6}$$

3.5. Learning Strategy

We adopt a learning strategy to realize two different modes in our SWARM. In one iteration step, we first use the clean loss to update the clean tokens and trigger. And then, we freeze the clean tokens and add the switch token to the input.

Figure 3. Visualization of clean and backdoor images.

We use backdoor loss and cross-mode feature distillation loss to update the switch token and trigger. Therefore, we need twice forward and backward propagations in one step to optimize the parameters.

4. Experiments

In this section, we will evaluate the performance of SWARM across various vision datasets, the impact of different hyper-parameters, and its robustness to backdoor detection and mitigation.

4.1. Experimental Setup

Datasets and models. We evaluate SWARM on datasets from the VTAB-1k [62] benchmark. Concretely, VTAB-1k is a collection of diverse visual classification tasks, which can be divided into three groups: Natural tasks contain natural images captured by standard cameras; Specialized tasks contain images captured by special equipment such as medical and satellite imagery; and Structured tasks require the models to have the geometric comprehension. In this collection, each task contains 1000 training samples and we use the provided split of the train set to evaluate our SWARM. In this scenario, only 800 samples are used for the training while the remaining 200 samples are used for validation. Besides, we select Vision Transformer (ViT) [15] which is pre-trained on Imagenet-21K [14] as the main target model.

Baselines and attack settings. We choose 4 existing backdoor attacks as our baselines: BadNets [25], Blended [13], WaNet [47] and ISSBA [37]. We adapt these attacks by setting the prompts as the only learnable parameters. We maintain the default settings following their papers to ensure the performance. We set the clean tokens to 50 in all cases and other settings of the visual prompt learning are drawn from [33]. For human imperceptibility, the ϵ is set to 4. The hyper-parameter λ is set to 100 as the default.

Evaluation metrics. There are two modes that need to be evaluated, where we use SWARM-C and SWARM-B to denote the model in the clean and backdoor mode, respectively. In clean mode, we should evaluate the model's performance under the clean images and images with triggers. Therefore, we use Benign Accuracy (BA) and Benign Accuracy with Triggers (BA-T) to measure the performance in the clean mode. While in the backdoor mode, we follow previous backdoor studies [22, 40, 47, 60], which use Benign Accuracy (BA) and Attack Success Rate (ASR) to measure the backdoor attack. Specifically, higher values of these metrics indicate the better performance.

4.2. Main Results

In this section, we perform our SWARM on VTAB-1k and present the results in Tab. 1.

SWARM-C correctly classifies clean images and triggered images. As observed in Tab. 1, SWARM-C can achieve comparable performance on both clean images and triggered images compared to the no-attack situation among all the datasets. In most cases, SWARM performs a minor accuracy drop of less than 2%. Meanwhile, no performance decline exists between triggered and clean images, indicating that even though the input images are triggered, it is difficult for the victims to detect performance differences under this mode. In some cases, SWARM-C outperforms the no-attack situation, ensuring its competitiveness.

SWARM-B correctly classifies clean images. After the switch token is added, the model is changed to the backdoor mode. In this situation, SWARM acts as a normal backdoor model which will manipulate the prediction result whenever there are backdoor triggers. As is shown in Tab. 1, SWARM-B achieves the best benign accuracy among all backdoor attacks, which has less than a 2% drop compared to the no-attack situation in most cases. The average of the SWARM's BA is also the highest in these methods.

SWARM-B achieves high attack success rates. We can see from Tab. 1 that SWARM shows promising performance in terms of ASR. Specifically, SWARM achieves high ASRs ($> 95\%$) on all datasets and a 97.90% on average. Moreover, 97.90% is the highest average ASR value among all backdoor attacks. Although ISSBA can achieve a very competitive ASR in some cases, it has some bad performance on specific datasets with less than 90% ASR.

4.3. Ablation Study

SWARM on different backbones. In this part, we evaluate our methods on different backbones. Visual prompts can not only be added to ViT but also they can be used for other backbones. As shown in Table 2, we further experiment with our SWARM on Swin Transformer [42] and ConvNeXt [43]. These pre-trained models are all pre-trained on the Imagenet-21K [14] and then adapted with our SWARM. Our method still has $96\%+$ ASRs for these two backbones. It demonstrates SWARM's effectiveness with no regard to the upstream backbones.

Effect of the number of switch tokens. In our SWARM, we adopt a single token as the switch. Here, we investigate the impact of varying the number of switch tokens on

Table 1. The main results (%) on VTAB-1k [62] dataset collection. SWARM is competitive with four advanced backdoor attacks in terms of BA, and meanwhile reaches high ASRs which exceeds 95%. We have marked the best BA and ASR on 5 backdoor attacks with bold scores while underlined scores are the second-best performance.

Attack→	No Attack	BadNets		Blended		WaNet		ISSBA		SWARM-B		SWARM-C	
Datasets-VTAB↓, Metric→	BA	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	BA-T
CIFAR-100	77.27	67.57	86.07	64.82	85.65	65.72	83.72	72.87	**99.28**	**76.36**	96.96	76.41	76.38
Caltech101	83.89	46.11	50.93	41.51	54.77	52.98	48.33	79.85	89.99	**82.63**	**96.58**	84.32	84.01
DTD	65.90	37.23	73.94	34.10	60.85	35.32	62.29	20.53	87.82	**62.11**	**95.11**	63.67	63.99
Flowers102	97.48	**94.73**	91.15	91.61	80.01	80.40	28.17	84.23	88.55	93.53	**96.99**	96.80	96.93
Pets	87.52	81.49	87.52	81.90	79.56	73.86	34.94	73.67	87.46	**86.02**	**98.53**	86.64	86.43
SVHN	68.76	61.39	90.04	62.83	91.79	50.58	33.09	66.63	**99.24**	**67.72**	96.05	67.84	68.81
Sun397	47.83	29.35	73.92	26.02	57.03	24.92	71.14	35.76	92.81	**43.53**	**96.53**	47.41	45.40
Patch Camelyon	75.01	69.62	70.63	67.15	75.73	63.62	82.71	72.98	96.43	**76.65**	**96.56**	78.37	77.83
EuroSAT	92.96	90.74	98.96	90.37	95.89	77.17	27.72	91.24	**99.67**	**91.94**	96.52	92.09	91.30
Clevr/count	45.73	42.36	**100.00**	42.77	**100.00**	38.67	96.19	43.70	**100.00**	**44.83**	99.98	45.60	45.53
Clevr/distance	54.13	**53.89**	99.98	51.39	**100.00**	40.75	64.23	52.26	**100.00**	49.37	99.99	50.98	50.37
DMLab	36.92	34.04	99.51	34.41	99.48	33.87	75.70	34.18	**99.56**	**34.34**	97.39	34.97	34.77
KITTI	66.38	60.90	**99.72**	62.59	96.06	63.71	92.12	64.70	96.77	**65.96**	98.87	69.20	62.59
dSprites/location	70.78	62.23	**100.00**	63.80	99.96	53.12	24.92	68.57	99.84	**68.83**	99.79	69.97	69.29
dSprites/orientation	35.39	26.27	**99.94**	29.55	99.87	24.91	48.62	33.82	99.83	**36.58**	99.62	36.39	36.41
SmallNORB/azimuth	11.96	9.31	96.40	7.65	79.25	7.72	77.23	**13.42**	**100.00**	9.95	99.06	13.55	13.43
SmallNORB/elevation	27.29	26.16	86.36	27.85	85.08	22.05	47.41	30.20	**99.89**	**30.77**	99.79	31.36	30.49
Average	61.48	52.55	88.53	51.78	84.76	47.61	58.74	55.21	96.32	**59.95**	**97.90**	61.50	60.82

SWARM-B: The switch token is added and the model is under backdoor mode.

SWARM-C: The switch token is removed and the model is under clean mode.Therefore, the images with triggers are still have normal performance.

Table 2. Results of SWARM on different backbones. It has the same performance comparing to the ViT.

Attack→	No Attack	SWARM-B		SWARM-C	
Backbones↓, Metric→	BA	BA	ASR	BA	BA-T
ViT	77.27	76.36	96.96	76.41	76.38
Swin-B	72.62	70.11	97.66	71.11	70.72
ConvNeXt-Base	73.31	73.43	96.64	75.51	76.24

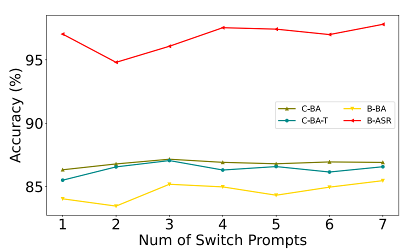


Figure 4. The effect of increasing the numbers of switch tokens.

three different datasets: CIFAR100 [34], Flowers [49], and Pets [50]. As demonstrated in Fig. 4, performance does not improve with an increasing number of switch tokens, suggesting that one switch token suffices for our method.

Effect of λ. In our SWARM, the weight of cross-mode feature distillation loss λ is set to 100. We further explore the impact of λ on the three previously mentioned datasets. As depicted in Fig. 5, performance remains robust when λ does not exceed 200, with 100 being the most suitable trade-off parameter.

Effect of the switch token and the cross-mode distillation loss. As illustrated in Tab. 3, we demonstrate the indispensability of each component. Without the switch token, SWARM exhibits poor performance across both modes. In the absence of cross-mode feature distillation loss, SWARM behaves normally in clean mode. However, the model cannot accurately classify clean images, resulting in a 10% drop in terms of BA in backdoor mode. Besides, we perform the ablation on trigger learning in Appendix.

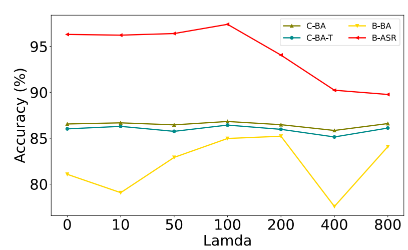
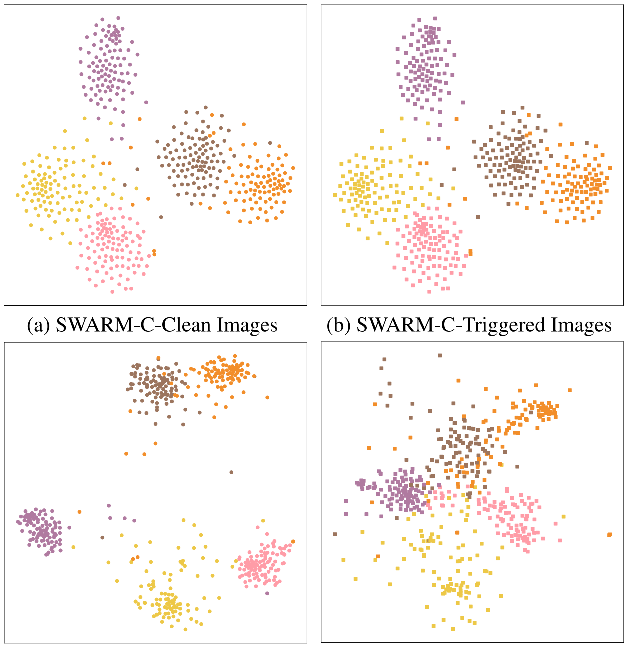


Figure 5. The effect of increasing the λ.

Table 3. Effect of the switch token S and the cross-mode distillation loss $\mathcal{L}_{cs}$ on three datasets.

Datasets	Mode	SWARM-B		SWARM-C	
/	Metric	BA	ASR	BA	BA-T
CIFAR100	w/o S	36.64	66.38	36.64	25.91
	w/o $\mathcal{L}_{cs}$	69.75	98.09	76.03	74.91
	w/ all	76.36	96.96	76.41	76.38
Flowers	w/o S	80.18	70.28	80.18	23.52
	w/o $\mathcal{L}_{cs}$	91.09	95.33	91.09	95.33
	w/ all	93.53	96.99	96.80	96.93
Pets	w/o S	76.45	68.68	76.45	25.92
	w/o $\mathcal{L}_{cs}$	82.37	95.50	87.19	86.92
	w/ all	86.02	98.53	86.64	86.43



(a) SWARM-C-Clean Images (b) SWARM-C-Triggered Images

(c) SWARM-B-Clean Images (d) SWARM-B-Triggered Images

Figure 6. The t-SNE visualization of features extracted by SWARM. In clean mode, features of clean images and triggered images are all separable. In backdoor mode, features of clean images are separable while these of triggered images cluster together.

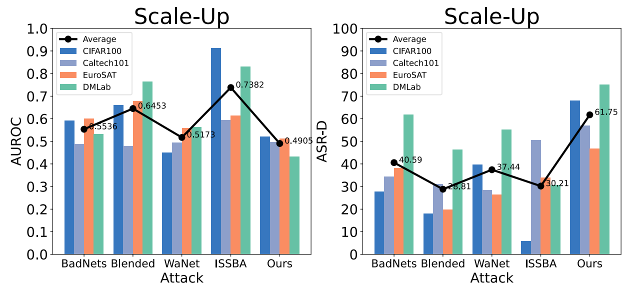


Figure 7. The results of Scale-Up detection method on five backdoor attacks. Lower AUROC and higher ASR-D indicates a better attack performance. Among these attacks, SWARM exceeds all other baseline attacks.

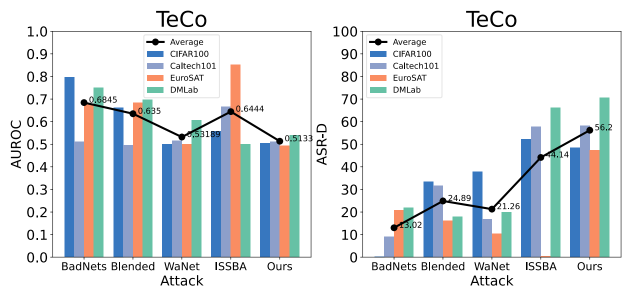


Figure 8. The results of TeCo detection methods on five backdoor attacks. Lower AUROC and higher ASR-D indicates a better attack performance. Among these attacks, SWARM exceeds all other baseline attacks.

4.4. Visualization

As shown in Fig. 6, we visualize the features obtained from the SWARM model. In the clean mode, we can observe that the clean features and the triggered have almost the same pattern and they are all separable, which explains the clean performance on the triggered images. In the backdoor mode, clean images' features are still separable which indicates the good prediction results on benign accuracy. In contrast, for triggered images' features in the backdoor mode, the situation is poles apart, i.e., the borders of the features are not as clear as the clean ones. The triggered images gather together so the classifier naturally makes the target predictions on these inputs. This phenomenon demonstrates our method's rationality and it is a worthy point to be further researched that only a small portion of parameters in the switch token can achieve such an obvious change in behaviors during the adaption.

4.5. Robustness to Backdoor Detection

In this section, we choose three backdoor detection methods[22] to check the attacks' stealthiness. They are Scale-Up [27], TeCo [41] and STRIP [23] (the results of STRIP are shown in Appendix). Following existing detection-based backdoor defenses [24, 26, 32], we use Area Under Receiver Operating Curve (AUROC) [17] as the metric, which is widely used to measure the trade-off between the false positive rate for clean samples and true positive rate for triggered samples. Besides, we adopt another metric called ASR-D. To calculate it, we first select out the clean samples categorized by the detection method and then use these samples to calculate the Attack Success Rate on the backdoored model. The higher ASR-D, the better the stealthiness of the backdoor attack since it means that detection methods can not detect the triggered samples. According to our threat model, we set the SWARM to clean mode when we meet the detection method and we test the attack performance of our method in the backdoor mode.

Scale-Up. Scale-Up [27] is a black-box input-level backdoor detection that only requires the predicted labels to detect the existence of backdoors. It discovers the phenomenon that the predictions of poisoned samples are significantly more consistent compared to those of benign ones when amplifying all pixel values. Based on this point, it evaluates the scaled prediction consistency between the input images and the scaled images and determines the malicious images based on a defender-specified threshold.

We use 3000 clean images and 3000 triggered images as the test set. We evaluate our method and all other baseline methods with AUROC and ASR-D. As shown in Fig. 7, we test the attack with the detection method on four different datasets which have covered all groups of datasets in VTAB-1K. They are CIFAR100 [34], Caltech101 [18], EuroSAT [29], and DMLab [8]. They are representatives

Table 4. The defense results on NAD. Our method still keeps high ASRs after the mitigation comparing to other baselines.

Attack→	BadNets		Blended		WaNet		ISSBA		Ours	
Dataset↓, Metric→	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
CIFAR100	73.84	57.54	72.24	61.59	69.87	6.69	81.92	5.21	75.80	98.92
Caltech101	81.44	10.09	82.83	8.00	82.46	9.33	91.72	1.23	81.75	97.15
EuroSAT	90.77	64.59	90.93	71.10	90.43	10.69	94.30	15.76	90.82	96.43
DMLab	34.03	32.43	34.63	25.37	52.24	30.99	53.54	24.29	33.24	99.15

Table 5. The defense results on I-BAU. Our method still keeps high ASRs after the mitigation comparing to other baselines.

Attack→	BadNets		Blended		WaNet		ISSBA		Ours	
Dataset↓, Metric→	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
CIFAR100	71.12	69.79	69.59	68.94	64.3	14.24	75.87	9.83	74.79	97.56
Caltech101	78.98	9.25	81.73	6.74	80.81	8.51	86.23	2.79	79.28	99.65
EuroSAT	92.07	85.11	92.07	85.11	86.41	13.57	92.95	16.26	86.74	99.77
DMLab	37.05	64.77	36.58	75.65	36.24	15.83	38.98	24.9	25.32	99.9

of natural tasks, specialized tasks, and structured tasks. In Fig. 7, we mark out all performance on four datasets and compare our method to the four baseline attacks. At the same time, we also calculate the average values to make a comprehensive comparison. As we can see in Fig. 7, on average, our method has the lowest AUROC (0.4905) and the highest ASR-D (61.75%) among these baseline attacks. Although WaNet has the same performance in AUROC, it shows a poor ASR-D. SWARM has the best ASR-D which is more than 20% higher than other baselines, indicating that our methods are undetectable.

TeCo. TeCo [41] is a test-time backdoor detection method that only uses the predicted labels to determine whether the model is backdoored. The main idea of this method is that backdoor-infected models have similar performance across different image corruptions on clean images and perform discrepantly on poisoned images. It first corrupts images with various corruption methods and different levels of severity, then uses this corruption set to evaluate the robustness consistency, and finally gets the result of detection.

Under this detection method, we also follow the Scale-Up detection settings, with 3000 clean samples and 3000 poisoned samples on four various datasets. The results are shown in Fig. 8. We can see that in most cases, SWARM has the lowest AUROC and the highest ASR-D among different attack methods against Scale-Up. Our method has an improvement of 12% in terms of ASR-D compared to ISSBA. Our experiments show that our SWARM is successful in overcoming the limitation of these baseline attacks, whose backdoor behavior is hard to be detected by these detection methods due to the switchable mechanism.

4.6. Robustness to Backdoor Mitigation

The users can employ extra clean data for backdoor mitigation [35, 38, 54, 56, 59, 60] to alleviate the backdoor threat. In this part, we allow users to tune the prompts' parameters while keeping the backbone frozen. Furthermore, we utilize extra 1000 clean test samples as the provided clean subset for mitigation.

NAD. Neural Attention Distillation (NAD) [39] is a backdoor mitigation method that employs a teacher network trained on a small clean data subset to guide the fine-tuning of the backdoored student network, ensuring alignment of intermediate-layer attention. As demonstrated in Tab. 4, we conduct experiments to mitigate model backdoors using the same settings provided in its original paper [39]. Due to the pre-trained scenario, we only apply the input embedding layer for neural attention distillation. As a result, our method maintains a high ASR under backdoor mitigation, exceeding 96% in all cases, while other baseline attacks exhibit lower ASRs. These results confirm that our method can resist NAD successfully.

I-BAU. I-BAU [61] is another backdoor mitigation method that leverages implicit hypergradient to account for the interdependence between inner and outer optimization, subsequently solving the min-max problem on clean data for unseen test data. Following the same settings in its original paper [61], we evaluate our SWARM and baselines on four datasets. As illustrated in Tab. 5, our method maintains over 97% ASRs, surpassing all other baseline attacks, while the ASRs of ISSBA and WaNet are lower than 25%, indicating a substantial performance gap.

All the above experiments on defense methods have demonstrated our method can resist the backdoor detection and backdoor mitigation, which further increases the risk of the proposed backdoor attack for the victims.

5. Conclusion

In this paper, we explored backdoor attacks on adapting pre-trained vision transformers to the downstream visual recognition tasks and identified a novel security threat towards such a paradigm. To be specific, we utilized an extra prompt token to toggle the backdoor model on or off. We then optimized trigger and prompt tokens with a clean loss, a backdoor loss, and a cross-mode feature distillation loss to achieve this mechanism. We showed that SWARM can maintain the accuracy of clean images compared to other methods while achieving high ASRs. Moreover, we demonstrated SWARM has a good ability to be undetected and can not be removed through backdoor defenses. To the best of our knowledge, we are the first to propose the switchable backdoor mechanism and tailor this kind of backdoor attack based on visual prompting. We hope that our work opens a new domain of attack mechanisms on pre-trained models, and can encourage future defense research.

Acknowledgement. This work is supported in part by the National Natural Science Foundation of China under Grant 62171248, Shenzhen Science and Technology Program (JCYJ20220818101012025), and the PCNL KEY project (PCL2023AS6-1).

References

[1] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1505–1521, 2021. 3

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[3] Jiawang Bai, Baoyuan Wu, Yong Zhang, Yiming Li, Zhifeng Li, and Shu-Tao Xia. Targeted attack against deep neural networks via flipping limited weight bits. *arXiv preprint arXiv:2102.10496*, 2021. 2

[4] Jiawang Bai, Kuofeng Gao, Dihong Gong, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Hardly perceptible trojan attack against neural networks with bit flips. In *ECCV*, 2022. 2

[5] Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *ECCV*, 2022. 1

[6] Jiawang Bai, Baoyuan Wu, Zhifeng Li, and Shu-Tao Xia. Versatile weight attack via flipping limited bits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[7] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 3

[8] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016. 7

[9] Nicholas Carlini. Poisoning the unlabeled dataset of {Semi-Supervised} learning. In *USENIX Security*, 2021. 2

[10] Aochuan Chen, Peter Lorenz, Yuguang Yao, Pin-Yu Chen, and Sijia Liu. Visual prompting for adversarial robustness. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 3

[11] Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *CVPR*, 2023. 1, 3

[12] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. 2022. 2

[13] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. 2017. 2, 5

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1, 5

[16] Wei Du, Yichun Zhao, Boqun Li, Gongshen Liu, and Shilin Wang. Ppt: Backdoor attacks on pre-trained models via poisoned prompt tuning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 680–686, 2022. 2

[17] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 7

[18] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 7

[19] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *AAAI*, 2023. 1, 3

[20] Kuofeng Gao, Jiawang Bai, Bin Chen, Dongxian Wu, and Shu-Tao Xia. Backdoor attack on hash-based image retrieval via clean-label data poisoning. In *BMVC*, 2023. 2

[21] Kuofeng Gao, Jiawang Bai, Baoyuan Wu, Mengxi Ya, and Shu-Tao Xia. Imperceptible and robust backdoor attack in 3d point cloud. *IEEE Transactions on Information Forensics and Security*, 19:1267–1282, 2023. 2

[22] Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor defense via adaptively splitting poisoned dataset. In *CVPR*, 2023. 5, 7

[23] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125, 2019. 7

[24] Yansong Gao, Yeonjae Kim, Bao Gia Doan, Zhi Zhang, Gongxuan Zhang, Surya Nepal, Damith C Ranasinghe, and Hyoungshick Kim. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *IEEE Transactions on Dependable and Secure Computing*, 19(4): 2349–2364, 2021. 7

[25] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *IEEE Access*, 2019. 2, 5

[26] Junfeng Guo, Ang Li, and Cong Liu. Aeva: Black-box backdoor detection using adversarial extreme value analysis. *arXiv preprint arXiv:2110.14880*, 2021. 7

[27] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023. 7

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[29] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 7

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[31] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *CVPR*, 2023. 1, 3

[32] Xijie Huang, Moustafa Alzantot, and Mani Srivastava. Neuroninspect: Detecting backdoors in neural networks via output explanations. *arXiv preprint arXiv:1911.07399*, 2019. 7

[33] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1, 3, 5

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 7

[35] Boheng Li, Yishuo Cai, Haowei Li, Feng Xue, Zhifeng Li, and Yiming Li. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In *CVPR*, 2024. 8

[36] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. In *NeurIPS*, 2024. 2

[37] 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*, pages 16463–16472, 2021. 5

[38] 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*, 34:14900–14912, 2021. 8

[39] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021. 8

[40] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2, 3, 5

[41] Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16363–16372, 2023. 7, 8

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 5

[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[44] Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Beyond sole strength: Customized ensembles for generalized vision-language models. *arXiv preprint arXiv:2311.17091*, 2023. 2

[45] Peizhuo Lv, Hualong Ma, Jiachen Zhou, Ruigang Liang, Kai Chen, Shengzhi Zhang, and Yunfei Yang. Dbia: Data-free backdoor injection attack against transformer networks. *arXiv preprint arXiv:2111.11870*, 2021. 2

[46] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2022. 1, 3

[47] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 5

[48] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020. 3

[49] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[50] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[52] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *CVPR*, 2022. 2

[53] Akshayvarun Subramanya, Aniruddha Saha, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Backdoor attacks on vision transformers. *arXiv preprint arXiv:2206.08477*, 2022. 2

[54] Ruixiang Tang, Jiayi Yuan, Yiming Li, Zirui Liu, Rui Chen, and Xia Hu. Setting the trap: Capturing and defeating backdoor threats in plms through honeypots. In *NeurIPS*, 2023. 8

[55] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 3

[56] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 8

[57] Junyang Wu, Xianhang Li, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Unleashing the power of visual prompting at the pixel level. *arXiv preprint arXiv:2212.10556*, 2022. 1, 3

[58] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 1

[59] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *ICLR*, 2024. 8

[60] Sheng Yang, Yiming Li, Yong Jiang, and Shu-Tao Xia. Backdoor defense via suppressing model shortcuts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 5, 8

[61] Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021. 8

[62] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1, 5, 6

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[64] Mengxin Zheng, Qian Lou, and Lei Jiang. Trojvit: Trojan insertion in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4025–4034, 2023. 2