Attention-Imperceptible Backdoor Attacks on Vision Transformers

Zhishen Wang[1,2], **Rui Wang**[1,2], **Lihua Jing**[1,2*]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
wangzhishen@iie.ac.cn, wangrui@iie.ac.cn, jinglihua@iie.ac.cn

Abstract

With the successful transition of Transformers from natural language processing (NLP) to computer vision (CV) domains, Vision Transformers (ViTs) have achieved state-of-the-art performance in many CV tasks. However, backdoor attacks, a significant threat in deep learning, also pose a risk to the security of ViT models. Recently, several backdoor attack methods targeting the patch-level self-attention mechanism in ViTs have been proposed, but they are relatively naive in terms of stealthiness and robustness against defensive measures, lacking in-depth investigation. In this paper, we explore the crucial role of attention-level imperceptibility in backdoor attacks for ViTs and propose an Attention-Imperceptible Backdoor Attacks on Vision Transformers (AIBA). In AIBA, a constrained adversarial perturbation is used as the trigger to achieve visual imperceptibility. Additionally, the trigger is designed to seamlessly implant into the focal areas of the image, ensuring that the trigger receives enough attention from the model without causing anomalies at the attention level. During the backdoor learning process, we designed an efficient constrained bi-level optimization training strategy at the mini-batch level to implant an effective backdoor in the victim model using the imperceptible trigger. We evaluated the effectiveness of the proposed AIBA across multiple datasets and ViT benchmarks and explored the robustness of AIBA against current ViT-specific defense methods. The experimental results demonstrate that our backdoor attack method can successfully implant a powerful and stealthy backdoor into ViTs.

1 Introduction

Vision Transformers (ViTs) (Dosovitskiy et al. 2020; Touvron et al. 2021; Mehta and Rastegari 2021; Liu et al. 2021; Yuan et al. 2021; Carion et al. 2020; Xie et al. 2021), by utilizing their advanced self-attention mechanisms to interpret images, have demonstrated performance that matches or even exceeds that of CNN architectures in a variety of computer vision tasks. Unfortunately, similar to CNNs, ViTs are also vulnerable to a variety of threats in deep learning, with backdoor attacks (Gu, Dolan-Gavitt, and Garg 2017; Liu et al. 2018; Nguyen and Tran 2020; Saha, Subramanya, and Pirsiavash 2020; Liu et al. 2020; Doan et al. 2021; Zhao et al. 2022; Nguyen and Tran 2021; Zhong, Qian, and Zhang 2022; Wenger et al. 2021) being one of the most dangerous. In a backdoor attack, an attacker poisons the trainng data and implants a backdoor during the model's training phase. The backdoor is triggered by specific inputs, allowing the victim model to perform normally on benign samples while performing predefined malicious behavior on backdoor samples. Due to the extensive parameters and complex training process of ViTs, users often rely on "outsourced" training or exploit pre-trained models from open-source communities to fine-tune the downstream tasks. These approaches provide attackers with more opportunities to implant backdoors, motivating a growing number of researchers to explore the security of ViTs under backdoor attacks.

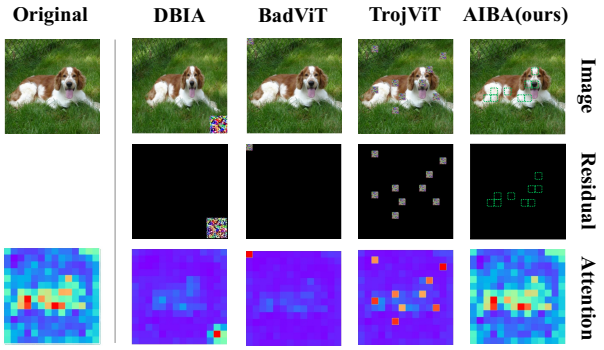

Figure 1: Visualization of backdoor images, residual maps and attention maps of the victim ViT models after different ViT-sepcefic attacks.

Recently, several backdoor attack methods designed specifically for ViT models have been proposed (Lv et al. 2021; Subramanya et al. 2022, 2024; Zheng, Lou, and Jiang 2023; Yuan et al. 2023). As is known, for a backdoor to be effective, it must maintain a high attack success rate while enhancing its stealthiness to better circumvent adversarial defenses. However, the previous attack methods designed for Vision Transformers are relatively naive, both in terms of the stealth and the robustness against defenses. Firstly, most existing backdoor attack methods targeting ViTs depend on visually conspicuous and distinct triggers. Without these triggers, the victim model either experiences a sig-

*Corresponding author

nificant drop in accuracy on clean inputs or a reduced attack success rate on backdoor samples. However, these distinct backdoor samples are easily filtered out during manual inspection, resulting in the failure of the attack. Additionally, all current backdoor attack methods for ViTs attempt to make the trigger capture the maximum attention of the model, which results in significant anomalies in the attention maps. Due to the self-attention mechanism, the ViTs are susceptible to interpretability algorithms that highlight the location of the trigger in test images, making such attention anomalies relatively easy to detect. Recent successful backdoor defense methods in ViTs (Subramanya et al. 2024; Doan et al. 2023) are also based on detecting anomalous attention values. Therefore, if the anomalies in the attention maps are not properly addressed, backdoor attacks stand very little chance to succeed.

To address these challenges, we propose AIBA, a novel, effective and stealthy backdoor attack mechanism specifically designed for ViTs. We first use the Attention-Max trigger selection method to identify the optimal trigger placement for each backdoor sample, making the attention to the trigger blend seamlessly into the original attention map. Then, we generate visually imperceptible backdoor triggers using adversarial perturbation techniques (Madry et al. 2017; Doan et al. 2021). During the trigger generation process, unlike previous ViT-specific backdoor attack methods that optimize the trigger to maximize the model's attention to it, we focus solely on establishing a backdoor path between the trigger and the target label and set it as our optimization objective so as to further reduce anomalies of the backdoor images at the attention level (various backdoor images and the attention map under different attacks are shown in Figure 1). Throughout the backdoor learning process, we define the trigger generation and model poisoning process as a constrained bi-level optimization problem. To effectively solve the optimization problem of embedding a strong backdoor in the victim model with subtle triggers, we design an efficient alternating optimizing strategy at the mini-batch level, which ensures the generation of the optimal triggers while successfully activating the backdoor with minimal fine-tuning iterations.

We summarize the contributions of this paper as follows:

- We investigate the crucial role of attention-level imperceptibility in backdoor attacks against ViTs and designed an attention-imperceptible trigger generation method. By adding invisible perturbation triggers to original key areas of the backdoor images, we achieved natural backdoor embedding while suppressing anomalies at the attention level.
- We design an effective backdoor learning strategy that addresses the bi-level optimization problem of trigger generation and model poisoning through mini-batch level alternating training, enabling subtle triggers to activate a strong backdoor.
- We conduct comprehensive experiments, and the results show that our approach outperforms state-of-the-art general and ViT-specific backdoor attacks in both attack effectiveness and robustness against defenses.

2 Related Work

General backdoor attacks. Backdoor attacks have emerged as a significant threat to deep neural networks. BadNet (Gu, Dolan-Gavitt, and Garg 2017) is the first method proposed to carry out such attacks by introducing triggers and altering target labels to poison a subset of the training data. The poisoned data is then used to train the model. To enhance the stealthiness of backdoor attacks, some research has begun to focus on exploring invisible backdoor attack techniques. Chen et al. proposed a stealthy strategy (Chen et al. 2017) that generates poisoned images by blending backdoor triggers with benign images. In WaNet (Nguyen and Tran 2021),attackers employed warp-based triggers, which are more difficult for human inspection to detect. Refool (Liu et al. 2020) conducted backdoor attacks by leveraging reflection phenomena in images as invisible triggers. Some recent approaches (Li et al. 2020; Doan et al. 2021; Doan, Lao, and Li 2021; Zhao et al. 2022; Gao et al. 2024) have utilized adversarial perturbations with l_p-norm constraint to generate backdoor triggers. The aforementioned methods have demonstrated satisfied attack effectiveness and stealthiness in CNNs for visual classification tasks. However, general attacks are less effective against ViT models because of the unique characteristics of ViTs.

ViT-specific backdoor attacks and defences. To address the limitations of general backdoor attacks, some researchers have delved deeper into the vulnerabilities of ViTs to backdoor attacks. Lv et al. introduced a data-free backdoor attack for ViTs named DBIA (Lv et al. 2021). This method employs additional surrogate data to generate an optimized trigger and implant the backdoor into the model. As is known, in ViTs, an image is initially divided into multiple small patches, which are then flattened into a one-dimensional sequence of visual tokens. The self-attention mechanism is then used to capture the global dependencies within the image, facilitating effective image understanding. Leveraging this characteristic, Zheng et al. investigated the threats posed by area-wise and patch-wise triggers to ViT and introduced TrojViT (Zheng, Lou, and Jiang 2023), which involves implanting a patch-wise trigger to create a Trojan composed of vulnerable bits within the parameters of a ViT stored in DRAM memory. Similarly, Yuan et al. proposed BadViT (Yuan et al. 2023) , which also employs a patch-wise trigger to catch the model's maximum attention. In response, researchers have developed several defense methods specifically designed to protect ViTs from backdoor attacks. Subramanya et al. introduced BDVT (Subramanya et al. 2024), a defense mechanism that suppresses trigger activation by placing a black patch over the area with the highest attention score. Doan et al. investigated the sensitivity difference of ViTs between backdoor and clean samples to different patch processes to identify and filter out backdoor samples. Considering that these ViT-specific backdoor defense methods tend to exploit the model's abnormal perception of poisoned samples at the attention level to detect or filter out backdoors, we further propose an attention-imperceptible backdoor attack method to advance the offensive and defensive games in the backdoor security of ViTs.

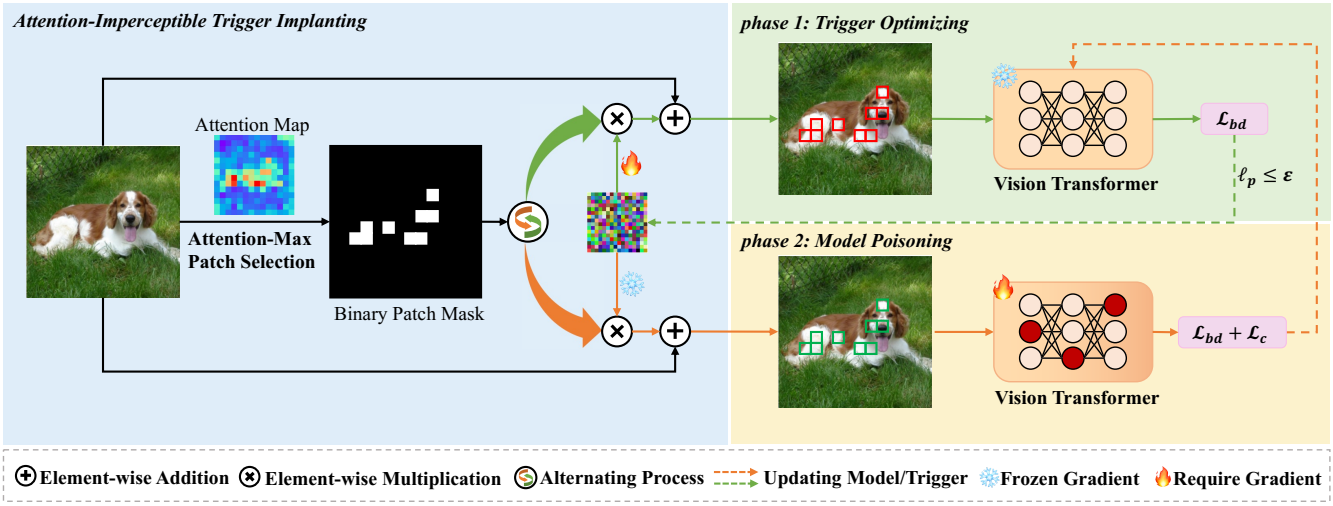

Figure 2: Overview of the proposed AIBA: AIBA begins by embedding an imperceptible trigger into the backdoor samples. During the backdoor learning process, it alternately optimizes the trigger and model parameters at the mini-batch level, enabling effective and stealthy backdoor attacks.

3 Methodology

3.1 Threat Model

We consider the same threat model as in prior studies (Nguyen and Tran 2021; Doan et al. 2021; Saha, Subramanya, and Pirsiavash 2020), assuming that the backdoor attack is performed during training, and the attacker has full access to the victim model, including the model's parameters and training process. The victim model is then uploaded to the internet and made publicly available, allowing users to download and deploy it in their applications.

3.2 Problem Formulation

Our approach focuses on backdoor attacks on ViTs in supervised classification tasks. A trained Vision Transformer classifier can be represented as $F_\theta : I \to L$, where the function maps the input image I to a classification label L. In this notation, θ represents the parameters of the ViT, $I \in \mathbb{R}^{C\times W\times H}$ is the input domain with C, H, W specifying the channel, height, and width of the image respectively. The output $L \in \mathbb{R}^k$ corresponds to the set of labels, where k is the number of classes. The ViTs split an input image into N_p patches, which are treated as visual words input to the Transformer blocks. Additionally, a classification token (CLS token) is added to the head of the patch sequence, forming the input token sequence for the Transformer as $T = \{t_{\text{cls}}, t_1, t_2, \ldots, t_{N_p}\}$. The token sequence is then used to calculate attention through the multi-head self-attention (MSA) modules as follows:

$$A(I) = Softmax\left(\frac{(TW^Q)(TW^K)}{\sqrt{d_k}}\right)(TW^V), \quad (1)$$

where the projection matrices $W^Q, W^K, W^V \in \mathbb{R}^{d_p\times d_k}$, $A(I) \in \mathbb{R}^{(N_p+1)\times d_k}$, d_p and d_k is the dimension of the flatten patches and the attention model.

After passing through multiple MSA blocks to compute attention, the output corresponding to the CLS token is used as a classification feature and fed into an MLP classifier to obtain the final class prediction $\hat{L}$, which can be denoted as:

$$\hat{L} = MLP(A(I)[0]) \quad (2)$$

To conduct the backdoor attack, we fine-tune the benign ViT model using a mixture of clean and poisoned samples. The benign model is initially trained on a clean dataset, denoted as $D_c = \{(x_i, y_i) : x_i \in I, y_i \in L, i = 0, 1, \ldots, N_d\}$. To create the poisoned subset D_b containing N_b images, we randomly select a proportion ρ of samples from the clean dataset and transform them into backdoor samples $(T_\phi(x), \eta(y))$. Here, $\rho = N_b/N_d$, referred to as the poisoning rate, is a hyperparameter. The function $T_\phi(\cdot)$ is used for backdoor injection, and $\eta(\cdot)$ alters the original labels to the target labels for the attack.

We adhere to the standard all-to-one backdoor objective, which seeks to establish a strong and covert backdoor between the backdoor samples $T_\phi(x)$ and the target labels $\eta(y)$, while maintaining normal performance on clean datasets. In other words, after fine-tuning the ViT classifier F_θ on a mixed dataset containing both backdoor samples and clean samples, we expect the behavior of F_θ to be altered as follows:

$$F_{\hat{\theta}}(x_i) = y_i, F_{\hat{\theta}}(T_\phi(x_j)) = \eta(y_j) \quad (3)$$

where $i \in [0, N_d - N_b], j \in [0, N_b]$ which indicate the indices of clean dataset and poisoned subset respectively.

For the target ViT model F_θ and the backdoor injection function T_ϕ, we minimize the following loss function to preserve performance on clean data while inducing the desired backdoor behavior:

$$\mathcal{L}_{clean} = \sum_{i=1}^{N_c} \mathcal{L}^{ce}(F_\theta(x_i), y_i), \quad (4)$$

$$\mathcal{L}_{backdoor} = \sum_{j=1}^{N_b} \mathcal{L}^{ce}(F_\theta(T_\phi(x_j)), \eta(y_j)), \tag{5}$$

where L^{ce} is the cross-entropy loss function, $N_c = N_d - N_b$ and N_b represent the number of clean and backdoor samples respectively, and $\frac{N_b}{N_d} = \rho$.

3.3 Attention-Imperceptible Trigger Generation

The success of a backdoor attack largely depends on the quality of the trigger generated by the backdoor injection function $T_\phi(\cdot)$. We expect a successful backdoor trigger for ViTs to possess the following characteristics: (1)Effectiveness: The generated trigger should be attached to the key areas of the image, ensuring that the model maximally focuses on the trigger so as to establish a strong backdoor pathway, thereby increasing the attack success rate. (2)Stealthiness: The trigger should be imperceptible both visually and attentionally. This means that the trigger should minimally alter the ViT model's original attention, ensuring that the areas of focus maintain consistency between clean and poisoned images. This helps to resist manual inspection and potential defense. To this end, we design Attention-Max Patch Selection and Adversarial Trigger Generation to achieve the effective and stealthy backdoor injection, as shown in Figure 2.

Attention-Max Patch Selection.

Given that ViT processes images as patch-wise visual tokens, we introduce our trigger at the patch level. Meanwhile, the choice of patch location greatly affects the success of the attack, so it is crucial to identify the patch positions that draw the most attention from the model.

Taking advantage of the Attention mechanism in ViTs, which inherently focuses on key areas of an image, we develope an Attention-Max Patch Selection method to determine the optimal locations for injecting the patch-level triggers. Given an input x, the ViT first generates an input token sequence $T = \{t_{\text{cls}}, t_1, t_2, \ldots, t_{N_p}\}$ through Patch Projection. After passing through multiple transformer blocks with multi-head attention computation, the CLS token, which guides the classification, encapsulates information about the attention allocated to different patches.

It is noteworthy that the embedding tokens obtained from input tokens after passing through multiple MSA computations have become increasingly mixed and lost the correlation information between the original input tokens, thereby diminishing the interpretability. Therefore, we use Attention Rollout (Abnar and Zuidema 2020) $\widetilde{A}^0_{l_{-1}}$ corresponding to the CLS token from the final transformer layer to identify the top-N patches that receive the most attention from the model. Then we generate a binary mask $M(x)$ to indicate the N_m patches of the input x that the model focuses on the most:

$$M(x) = \begin{cases} 1, & \text{if } \widetilde{A}^0_{l_{-1}}[i] \in \text{top}\left(\widetilde{A}^0_{l_{-1}}[1:N_p], N_m\right) \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where $\text{Top}(\cdot, N)$ denotes the selection of the to p patch positions with the highest attention values of the last ViT layer. In the subsequent backdoor injection process, we attach the patch-level triggers to these key positions, ensuring that our triggers receive enough attention from the victim ViT, thereby increasing the attack success rate. Moreover, attaching the trigger within these attention-max patches helps minimize anomalies in the attention map, making the attack imperceptible at the attention level.

Adversarial Trigger Generation.

Inspired by adversarial learning methods (Goodfellow et al. 2014; Madry et al. 2017; Arjovsky, Chintala, and Bottou 2017), we designe a patch-level trigger generation method based on adversarial perturbations to achieve visual imperceptibility. We start by randomly initializing a patch-level perturbation $\phi \in \mathbb{R}^{\text{ps}\times\text{ps}\times C}$, where ps is the patch size as set in the targeted ViT, and C is the number of channels in the input image x. We then attach the perturbation-based trigger to the input image x according to the binary mask $M(x)$:

$$T_\phi(x) = x + M(x) \cdot \phi \tag{7}$$

In previous works, the trigger generation methods for ViT optimized the trigger to capture the model's maximum attention. However, we find that this optimization strategy might alter the model's original attention, causing anomalies of the backdoor samples at the attention level. Therefore, we directly attach the trigger to the patches where the model naturally has the highest attention. This ensures that the model maximally focuses on the triggers without altering its attention to the backdoor images. Since our trigger is strategically placed in the most attention-drawing patches, we only need to optimize the mapping of the backdoor samples to the target label. The optimization process of the trigger can thus be expressed as:

$$\hat{\phi} = clip_\epsilon(\phi - lr_p \cdot \nabla_\phi \mathcal{L}_{backdoor}), \tag{8}$$

where ϕ is the perturbation-based trigger, lr_p is the learning rate for trigger optimization and $clip_\epsilon$ is the clip function to constrain each pixel of the generated triggers to be within an ϵ neighborhood of the corresponding pixel in the input image x, ensuring the visual imperceptibility of the backdoor samples.

3.4 Effective Backdoor Learning

To create a strong backdoor pathway in the poisoned model using a subtle perturbation-based trigger while maintaining the model's accuracy on benign samples, we have introduced our effective backdoor learning process as a bi-level constrained optimization problem:

$$\begin{aligned} \min_{\theta} \quad & \alpha\mathcal{L}_{clean}(\theta) + \beta\mathcal{L}_{backdoor}(\theta, \hat{\phi}) \\ s.t. \quad & (i) \quad \hat{\phi} = \arg\min_{\phi} \mathcal{L}_{backdoor}(\hat{\theta}, \phi) \\ & (ii) \quad d(T_\phi(x), x) \leq \epsilon \end{aligned} \tag{9}$$

Where α and β determine the weighting of the loss contributions from the clean and backdoor data during the tuning of the ViTs and $d(\cdot) = \|\cdot\|_p$ represents the l_p-norm distance in the pixel space. In this bi-level optimization problem, our

objective is to simultaneously optimize the backdoor injection function T_ϕ, with parameters ϕ, and the victim model f_θ, with parameters θ. We aim to find an optimal solution $(\hat{\theta}, \hat{\phi})$ so that the victim classification model $f_{\hat{\theta}}$ can misclassify the backdoor-injected samples $T_{\hat{\phi}}(x)$ into the target category $\eta(y)$ while accurately classifying the clean samples.

The optimization problem presented in Equation 9 is challenging. First, because our triggers are in the form of subtle perturbations, backdoor samples containing these triggers differ minimally from clean samples. From the model's perspective, these poisoned samples resemble difficult-to-classify points near the decision boundary. Consequently, fine-tuning the model on such backdoor samples may shift the decision boundary, resulting in reduced accuracy on benign samples. Second, optimizing the backdoor injection function T_ϕ is a non-convex problem. If trigger generation and model poisoning are conducted in separate training stages, the trigger optimization process is more likely to become trapped in a bad local optimum, making it difficult to execute a successful attack.

To overcome these challenges, we designed an alternating update strategy that simultaneously optimizes the trigger ϕ and the victim model parameters θ during backdoor learning. Taking into account the high computational cost of training ViTs, we propose a novel alternating optimization strategy at the mini-batch level which can implant an effective backdoor with minimal fine-tuning iterations. First, within each mini-batch, we perform $N_{\text{injection}}$ iterations on the backdoor samples to update the trigger ϕ, ensuring it achieves optimal performance with the current model parameters θ. Afterward, we use the optimized trigger $\hat{\phi}$ to conduct poisoned training and update the model parameters θ. In the subsequent mini-batch, the updated model $\hat{\theta}$ is used to refine the trigger, and this process continues iteratively. With this alternating optimization strategy, we can implant a powerful backdoor into the model within just a few training iterations (fine-tuning for only one epoch in our experiments). The complete backdoor learning workflow is depicted in Figure 2.

4 Evaluation

4.1 Experimental Setup

Datasets and Models. To evaluate the effectiveness of the proposed AIBA, we conducted backdoor experiments using various ViT models on two common image classification datasets: ImageNet (Russakovsky et al. 2015) and CIFAR-10 (Krizhevsky, Hinton et al. 2009). ImageNet is a benchmark dataset in computer vision, comprising 1.28 million training images, 50,000 validation images across 1,000 class labels. CIFAR-10, on the other hand, includes a total of 60,000 color images categorized into 10 classes, with 50,000 designated for training and 10,000 for testing. Given the large number of parameters in ViT models, backdoor attacks on ViTs generally involve fine-tuning pre-trained models to implant the backdoor. Accordingly, we select the officially pre-trained DeiT and ViT models, along with their various parameter versions, as our benign models for visual classification tasks. It is important to note that all the official pre-trained models were originally trained on the ImageNet; therefore, the benign models for CIFAR-10 were fine-tuned from the official pre-trained models.

Attacks	Clean	Backdoor				
	CA	BA	ASR	ARES ↓	APSNR ↑	ALPIPS ↓
BadNet	81.80	80.42	96.03	0.091	19.32	0.0024
WaNet	81.80	80.33	93.84	**0.011**	16.26	0.035
DBIA	81.80	80.65	97.38	0.503	12.07	0.0081
TrojViT	81.80	81.22	98.98	0.399	12.63	0.022
BadViT	81.80	81.04	99.76	0.417	11.63	0.0057
AIBA	81.80	**81.54**	**99.97**	0.078	**20.68**	**0.000058**

Table 1: Attack Effectiveness and Stealthness of AIBA and prior backdoor attacks on Deit-base with ImageNet.

Evaluation settings. In our experiments, we resize the input images to dimensions of $3 \times 224 \times 224$ and set the patch size to 16, generating $N_p = 196$ patches. We construct our mixed poisoned dataset with a poisoning rate of $\rho = 0.1$, using label 1 as the target label. In the Attention-Imperceptible Trigger Generation process, we select the top $N_m = 24$ patches that capture the model's highest attention to embed the imperceptible trigger. During trigger generation, We use l_∞ to constrain the trigger and limite the perturbation range to $\epsilon = 4/255$ to maintain stealthiness. The learning rate for trigger optimization lr_p is set at 0.01 and the $N_{\text{injection}}$ is set to 1. In the model poisoning process, we finetune the ViTs for 1 epoch with a batch size of 64 and a learning rate of 1e-5. The trigger generation and model poisoning are alternately performed in different mini-batches to learn effective and stealthy backdoors. The specific algorithmic process for backdoor learning in AIBA is provided in the supplementary materials.

Evaluation Metrics. To evaluate the effectiveness and stealthiness of the proposed AIBA, we define the following evaluation metrics.

- Clean Accuracy (CA), the accuracy of the clean test datasets on clean models;
- Backdoor Accuracy (BA), the accuracy of benign test samples on backdoor models;
- Attack Success Rate (ASR), the proportion of attacked samples which are successfully predicted as the target label.
- Attention Residual (ARES): We defined

$$ARES = avg(\sum_{i=1}^{N_d}(\sum_{j=1}^{N_p}(|A_{ij}^b - A_{ij}^c|))), \tag{10}$$

  where N_d is the number of the validation dataset, N_p is the number of patches in ViTs, A^b and A^c indicate the attention map of the clean and backdoor images respectively.
- Attention PSNR (APSNR): The average PSNR (Huynh-Thu and Ghanbari 2008) between the attention maps of the clean images and the backdoor images.
- Attention LPIPS (ALPIPS): The LPIPS (Zhang et al. 2018) between the attention maps of the clean images and the backdoor images.

ViT Models	Clean	Backdoor				
	CA	BA	ASR	ARES↓	APSNR↑	ALPIPS↓
ViT-base	81.41	81.25	99.94	0.061	21.58	0.000048
DeiT-Tiny	72.12	71.70	99.96	0.063	22.27	0.000053
DeiT-Small	79.83	79.46	99.98	0.076	18.92	0.000079
DeiT-Base	81.80	81.54	99.97	0.078	20.68	0.000058

Table 2: The attack results of AIBA under ImageNet.

ViT Models	Clean	Backdoor				
	CA	BA	ASR	ARES↓	APSNR↑	ALPIPS↓
ViT-base	98.33	97.62	99.98	0.068	21.17	0.000033
DeiT-Tiny	96.71	96.74	99.99	0.054	21.83	0.000030
DeiT-Small	98.31	97.91	100.00	0.064	20.85	0.000062
DeiT-Base	98.67	98.60	99.99	0.038	24.65	0.000014

Table 3: The attack results of AIBA under CIFAR-10.

4.2 Attack Experiments

Comparing against prior backdoor attacks. We compared the performance of the proposed AIBA against state-of-the-art techniques in attacking the DeiT-Base model on the ImageNet dataset, as shown in Table 1. Our AIBA achieves the highest ASR of 99.97% while maintaining the highest BA of 81.54%, closely aligning with CA and resulting in only a 0.26% drop in accuracy on clean samples. In terms of stealthiness, owing to our seamless and natural trigger embedding strategy, AIBA achieves a low ARES of 0.078, a high APSNR of 20.68, and a low ALPIPS of 0.000058, indicating that our method effectively minimizes abnormal attention, thereby achieving attention-level imperceptibility. For comparison, we selected the perceptible BadNet and the imperceptible WaNet, two of the most representative general backdoor attacks and DBIA (Lv et al. 2021), TrojViT (Zheng, Lou, and Jiang 2023) and BadViT-Inv (Yuan et al. 2023), three ViT-specific backdoor attacks. Unfortunately, being not specifically tailored for ViTs, both BadNet and WaNet only achieved ASR of approximately 95% and resulted in approximately a 2% reduction in the accuracy of clean samples when attacking the ViTs. Furthermore, we found that BadNet and WaNet perform well on the proposed stealthiness metric (with WaNet even outperforming our proposed AIBA on ARES). This is because these attack methods do not interfere with the attention mechanisms of ViTs during training, resulting in a smaller average change in the attention map. Nevertheless, unavoidable anomalies still appear in their attention maps. We give examples in Figure 3 to compare the stealthiness at the attention level between BadNets, WaNet and our AIBA. The ViT-specific backdoor attacks demonstrate comparatively higher ASR and lower accuracy loss on clean samples compared than general attacks. However, they all optimize trigger by maximizing the model's attention, which significantly alters the model's attention maps, leading to attention anomalies. As a result, these methods exhibit relatively high ARES and low APSNR, indicating lower stealthiness at the attention level. In contrast, our AIBA not only achieves optimal ASR and BA but also effectively mitigates attention-level anomalies, enabling both effective and stealthy backdoor attacks.

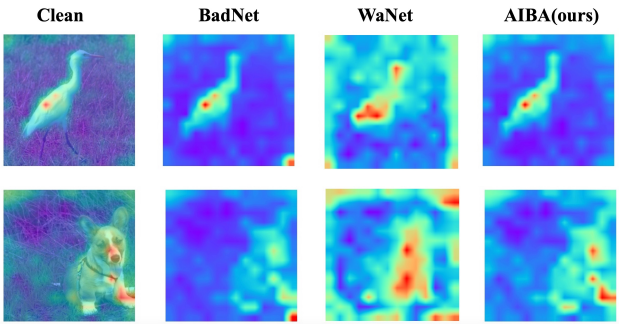

Figure 3: Examples of the attention map of the victim model under different attacks.

Performance in various ViTs and Datasets. To validate the effectiveness and stealthiness of our AIBA method across different datasets and ViT models, we applied AIBA to attack several ViT variants, including ViT-base (ViT-b), as well as DeiT-Tiny (DeiT-t), DeiT-small (DeiT-s), and DeiT-base (DeiT-b), using both the ImageNet and CIFAR10 datasets. The results are presented in Table 2 and Table 3. As shown, our AIBA method consistently achieved an attack success rate close to 100% across various datasets and models, while maintaining BAs close to CAs in each setting. We observed that on the large-scale ImageNet dataset, as the parameter size of the attacked model decreases, the difference between BA and CA becomes more pronounced, with the drop increasing from 0.26% for DeiT-base to 0.42% for DeiT-tiny. The reason is that in AIBA, the trigger is relatively subtle, leading to minimal differences between the backdoor samples and clean samples, and models with fewer parameters generally have simpler decision boundaries compared to those with larger parameters, making them more easily confused by these "hard" backdoor samples, resulting in a greater reduction in BA. However, the ASRs remain consistently high across all models, demonstrating that after applying our well-designed backdoor learning method, even these subtle triggers are sufficient to activate a powerful backdoor. When conducting backdoor attacks on the smaller CIFAR-10 dataset, the relatively simpler classification task allows our AIBA method to achieve ASRs near 100% and BAs closely aligned with CAs. In terms of stealthiness, our method consistently achieved an ARES below 0.1, an APSNR near 20 and an ALPIPS below 0.0001 when attacking different models on different datasets. In summary, our proposed AIBA method seamlessly integrates the trigger into the attacked images, achieving effective and successful backdoor attacks while maintaining stealthiness.

Influence of the trigger area in AIBA. The number of patches selected for embedding the trigger can significantly affect both the effectiveness and stealthiness of the backdoor attack. As shown in Table 4, the ASR tends to increase with the number of triggers, but beyond a certain point, adding more triggers no longer significantly improves ASR and instead reduces attention-level stealthiness. This occurs because when the number of triggers exceeds the image's pri-

Trigger Num	0	4	8	16	24	32
BA	81.80	81.55	81.57	81.50	81.54	81.57
ASR	0.01	99.76	99.80	99.93	99.97	99.98
ARES	0	0.047	0.068	0.072	0.078	0.081
APSNR	inf	27.24	23.73	22.18	20.68	19.74

Table 4: Influence of trigger area in AIBA.

ϵ of triggers	0	2/255	4/255	8/255	16/255	32/255
BA	81.80	81.52	81.54	81.57	81.54	81.52
ASR	0.01	99.82	99.97	99.98	99.99	99.99

Table 5: Stealthiness budget ϵ and attack effectiveness.

mary focus areas, the model inevitably extends its attention beyond the original regions to establish the backdoor connection between the trigger and the target label, resulting in higher ARES and lower APSNR. To balance attack effectiveness and stealthiness, we chose a trigger count of 24 as the default setting, where ARES remains below 0.1 and APSNR exceeds 20. This configuration enables our AIBA method to achieve an impressive 99.97% attack success rate while maintaining a high degree of stealthiness.

Influence of the Stealthiness Budget ϵ. We show the influence of ϵ on BA and ASR with DeiT-base in Table 5. The results show that as ϵ increases, the trigger becomes more conspicuous, resulting in a more effective backdoor attack. However, we are pleased to find that in AIBA, the trigger's prominence has little impact on BA. Moreover, even with a trigger perturbation range smaller than 2/255, our model still achieves an ASR of 99.82%. This indicates that the designed backdoor learning process can effectively utilize subtle triggers to establish a strong backdoor while minimizing the impact on the model's original performance on clean samples. Additionally, we showcase backdoor samples with triggers of varying saliency and explore the impact of using different l_p-norm constraints in supplement materials.

4.3 Defense Experiments

Here, we present the results of the proposed AIBA in resisting ViT-specific defense methods. Experiments of the performance against other general defense mechanisms are provided in the supplementary materials.

BDVT. This method defends against backdoor attacks by placing a black patch at the position with the highest heatmap score. To defence the multiple-patch attacks, we purposely enhanced BDVT by masking the top-N patch positions on the attention map. The results are presented in Table 6. It can be observed that as the number of masked patches increases, the ASR decreases, but the BA also suffers a proportional decline. Since the trigger is seamlessly integrated into the critical regions of the image in AIBA, and the difference in attention activation areas between clean and backdoor samples are difficult to distinguish, this straightforward defense method, which relies on detecting anomalies at the attention level to filter out backdoors, is ineffective against our AIBA attack.

Top N	0	1	2	4	8	16
BA	81.54	81.50	81.45	81.35	81.02	77.53
ASR	99.97	99.93	99.89	99.80	99.26	85.18

Table 6: The performance of AIBA against DBVT defense.

Process Ratio	k_d	k_s	TPR	TNR
0.1	20	0	2.4%	89.4%
0.2	37	0	3.8%	89.2%
0.4	57	0	2.6%	91.0%
0.8	69	0	1.7%	91.0%

Table 7: Defending of AIBA against PatchProcess, which tests TPR (%) and TNR (%) under different process ratio.

PatchProcess. This method comprises two methods, PatchDrop and PatchShuffle. We applied the two processes with varying ratios to a set of clean samples for $T = 100$ iterations and recorded the number of label changes to determine the thresholds k_d and k_s for each process. The results are summarized in Table 7. As the process ratio increases, the threshold k_d for PatchDrop also rises, indicating that clean samples are increasingly affected by PatchDrop. However, the threshold k_s for PatchShuffle remains at 0, suggesting that PatchShuffle does not work as expected. Therefore, we only analyse the results of PatchDrop. According to the PatchDrop theory, backdoor samples are more affected by PatchDrop compared to clean samples. Therefore, only samples with a number of changes exceeding k_d are classified as backdoor samples. Unfortunately, the TPR (True Positive Rate), which indicates detection success, remains very low across different process ratio, dropping below 4%. This suggests that this defense method is ineffective at detecting our backdoor samples. The is because through our effective backdoor learning process, the triggers gradually become characteristics of the target class and are naturally embedded in the key regions of the image, making the sensitivity difference to PatchDrop disappear.

The experiments have demonstrated that our AIBA exhibits strong stealth and robustness, effectively bypassing current defense mechanisms designed for ViTs.

5 Conclusion

In this paper, we explored the critical role of attention-level imperceptibility in backdoor attacaks on ViTs, and introduced a ViT-specific backdoor attack method AIBA. In AIBA, imperceptible triggers are seamlessly embedded into backdoor images, and an efficient mini-batch level alternating optimization strategy is employed to achieve an effective and stealthy backdoor. Experimental results demonstrate that the proposed AIBA achieves a near 100% attack success rate while maintaining a high level of stealth. We hope this work advances the development of offensive and defensive games in the backdoor security of ViTs and inspires the creation of more secure ViT models.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China Under Grants No. 62176253.

References

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Doan, K.; Lao, Y.; and Li, P. 2021. Backdoor attack with imperceptible input and latent modification. *Advances in Neural Information Processing Systems*, 34: 18944–18957.

Doan, K.; Lao, Y.; Zhao, W.; and Li, P. 2021. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11966–11976.

Doan, K. D.; Lao, Y.; Yang, P.; and Li, P. 2023. Defending backdoor attacks on vision transformer via patch processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 506–515.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, Y.; Li, Y.; Gong, X.; Li, Z.; Xia, S.-T.; and Wang, Q. 2024. Backdoor Attack with Sparse and Invisible Trigger. *IEEE Transactions on Information Forensics and Security*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

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

Li, S.; Xue, M.; Zhao, B. Z. H.; Zhu, H.; and Zhang, X. 2020. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5): 2088–2105.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.-C.; Zhai, J.; Wang, W.; and Zhang, X. 2018. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc.

Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 182–199. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lv, P.; Ma, H.; Zhou, J.; Liang, R.; Chen, K.; Zhang, S.; and Yang, Y. 2021. Dbia: Data-free backdoor injection attack against transformer networks. *arXiv preprint arXiv:2111.11870*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mehta, S.; and Rastegari, M. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

Nguyen, A.; and Tran, A. 2021. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*.

Nguyen, T. A.; and Tran, A. 2020. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33: 3454–3464.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2020. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11957–11965.

Subramanya, A.; Koohpayegani, S. A.; Saha, A.; Tejankar, A.; and Pirsiavash, H. 2024. A Closer Look at Robustness of Vision Transformers to Backdoor Attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3874–3883.

Subramanya, A.; Saha, A.; Koohpayegani, S. A.; Tejankar, A.; and Pirsiavash, H. 2022. Backdoor attacks on vision transformers. *arXiv preprint arXiv:2206.08477*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wenger, E.; Passananti, J.; Bhagoji, A. N.; Yao, Y.; Zheng, H.; and Zhao, B. Y. 2021. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6206–6215.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; and Wu, W. 2021. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 579–588.

Yuan, Z.; Zhou, P.; Zou, K.; and Cheng, Y. 2023. You Are Catching My Attention: Are Vision Transformers Bad Learners under Backdoor Attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24605–24615.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, Z.; Chen, X.; Xuan, Y.; Dong, Y.; Wang, D.; and Liang, K. 2022. Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15213–15222.

Zheng, M.; Lou, Q.; and Jiang, L. 2023. Trojvit: Trojan insertion in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4025–4034.

Zhong, N.; Qian, Z.; and Zhang, X. 2022. Imperceptible backdoor attack: From input space to feature representation. *arXiv preprint arXiv:2205.03190*.