

Summary of Changes

Dear Editors:

We would like to submit our manuscript “An Effective and Resilient Backdoor Attack Framework against Deep Neural Networks and Vision Transformers”, which we wish to be considered for publication in “IEEE Transactions on Dependable and Secure Computing”. This paper is an extended version of our previous paper, which is published in 2022 NDSS. We extend our previous work by extending the attack framework against vision transformers. While numerous studies have explored backdoor attacks against Convolutional Neural Networks (CNNs), there is a dearth of research on backdoor attacks tailored for vision transformers. Moreover, existing transformer backdoor attacks use visible triggers to launch the attacks [34], [63], making it easy for human defenders to detect abnormalities through visual inspections. By extending our advanced backdoor attack framework to ViT models, we aim to drive the advancement of backdoor attacks. In addition, we conduct experiments to compare our proposed method with state-of-the-art ViT backdoor attacks, including DBIA [34], DBAVT [9], and BAVT [49]. We also perform ablation studies to assess the effectiveness of different attack modules against ViT models. Furthermore, we demonstrate the resilience of our proposed attack against state-of-the-art ViT-specific backdoor defenses. In this paper, to enhance the attack effectiveness, we incorporate gradient enhancement techniques during trigger generation. Specifically, we assign higher weights to the gradients of selected neurons, which are crucial for classifying the target label. This prioritization allows us to amplify the poisoning effect of the generated trigger during the gradient descent optimization process.

ATTEQ-NN: Attention-based QoE-aware evasive backdoor attacks. Network and Distributed System Security Symposium, 2022.

Yours sincerely,

Xueluan Gong, Bowei Tian, Meng Xue, Yuan Wu, Yanjiao Chen, and Qian Wang

An Effective and Resilient Backdoor Attack Framework against Deep Neural Networks and Vision Transformers

Xueluan Gong, Bowei Tian, Meng Xue, Yuan Wu, Yanjiao Chen, *Senior Member, IEEE*, and Qian Wang, *Fellow, IEEE*

Abstract—Recent studies have revealed the vulnerability of Deep Neural Network (DNN) models to backdoor attacks. However, existing backdoor attacks arbitrarily set the trigger mask or use a randomly selected trigger which restricts the effectiveness and robustness of the generated backdoor triggers. In this paper, we propose a novel attention-based mask generation methodology that searches for the optimal trigger shape and location. To make the backdoored samples more natural, we introduce a Quality-of-Experience (QoE) term into the loss function and carefully adjust the transparency value of the trigger. To further improve the prediction accuracy of the victim model, we proposed an alternating retraining algorithm in the backdoor injection process. Besides, we launch the backdoor attack under a co-optimized attack framework that alternately optimizes the backdoor trigger and backdoored model to further improve the attack performance. Apart from DNN model, we also extend our proposed attack method against vision transformers. We evaluate our proposed method with extensive experiments on VGG-Flower, CIFAR-10, GTSRB, CIFAR-100, and ImageNette datasets. It is shown that we can increase the attack success rate by as much as 82% over baselines when the poison ratio is low and achieve a high QoE of the backdoored samples. Our proposed backdoor attack framework also showcases robustness against state-of-the-art backdoor defenses.

Index Terms—Backdoor attacks, Quality-of-Experience (QoE), attention mechanism, co-optimization framework.

1 INTRODUCTION

DEEP neural networks have made tremendous progress in past years and are applied to a variety of real-world applications, such as face recognition [47], automatic driving [37], natural language processing [36], and objective detection [44], due to superhuman performance. Vision transformer (ViT) [10] is a promising deep learning architecture that offers a compelling alternative to traditional convolutional neural networks (CNNs) for computer vision applications. Despite the success in the computer vision domain, both DNN and vision transformers are vulnerable to backdoor attacks [6], [17], [30], [34], [63]. It is shown that the attacker can inject a backdoor (a.k.a. trojan) into the model by poisoning the training dataset in the training time. The backdoored model behaves normally on the benign samples but predicts any sample attached with the backdoor trigger to a target false label. Due to its concealment, detecting backdoor attacks is very difficult. Moreover, the emergence of invisible backdoor triggers makes it more difficult to inspect whether the training samples are backdoored or not.

There exists a long line of backdoor attack strategies exploring

injecting backdoors into DNNs [17], [20], [21], [28], [30], [33], [45], [46], [56], [61]. However, they face the following shortcomings. First of all, most of the existing approaches [17], [33] use a random backdoor trigger or random trigger mask (random pattern and location) in the attack, which is easy to be detected and achieves a sub-optimal attack performance. Second, current backdoor attacks [17], [20], [21], [28], [33], [45], [46], [56] separate the trigger generation process from the backdoor injection process, thus resulting in generating sub-optimal backdoor trigger and backdoored model. Third, various works utilize visible backdoor triggers [5], [15], [17], [30], [33], [40], [43], which can be easily detected by visual inspection. Finally, although various existing works claimed to be defense-resistant [15], they can still be detected by the latest defenses, such as NAD [26] and MNTD [59]. In terms of backdoor attacks against ViTs, existing transformer backdoor attacks use visible triggers to launch the attacks [34], [63], making it easy for human defenders to detect abnormalities through visual inspections. Doan [9] proposed to generate hidden triggers based on a global warp of WaNet [45], but the attack success rate and the perceptual trigger quality are relatively low.

In this paper, we put forward a novel backdoor attack strategy that integrates effectiveness and evasiveness. From the attack effectiveness perspective, unlike the existing works that use fixed trigger masks (e.g., a square in the lower right corner), we utilize an attention map to differentiate the weights of the pixels. The mask is determined as the pixels with the highest weights since such pixels have a higher impact on the classification. Using such a carefully designed trigger mask, we can achieve a higher attack success rate than the existing works with the same trigger size. Moreover, rather than separating the backdoor trigger generation from the backdoor injection process, we adopt the co-optimization backdoor framework, that jointly optimizes the backdoor trigger

- X. Gong is with the School of Computer Science, Wuhan University, China. E-mail: xueluangong@whu.edu.cn.
- M. Xue is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China. E-mail: cseuxuemeng@ust.hk.
- Y. Wu is with the School of Computer Science and Artificial Intelligence & Engineering Research Center of Hubei Province for Clothing Information, Wuhan Textile University. E-mail: wuyuanxu@whu.edu.cn.
- Y. Chen is with the College of Electrical Engineering, Zhejiang University, China. E-mail: chenyanjiao@zju.edu.cn.
- B. Tian and Q. Wang are with the School of Cyber Science and Engineering, Wuhan University, China. E-mail: boweitian@whu.edu.cn, qianwang@whu.edu.cn.
- The first two authors have equal contributions.

and the backdoored model to generate an optimal backdoor trigger and achieve a higher attack success rate. In terms of the evasiveness goal, it is quantified by both the human vision and state-of-the-art defense strategies. We carefully adjust the transparency (i.e., opacity) of the backdoor trigger and add a Quality-of-Experience (QoE) constraint to the loss function, aiming to generate a more natural backdoor trigger. Furthermore, we propose an alternating retraining algorithm that updates the model using either mixed samples or only clean samples according to the iteration index. In addition to DNN models, we extend our proposed attack method to target vision transformers. Experiments show that our proposed method outperforms baselines in both attack success rate and clean data accuracy, especially when the poison ratio is low. It is demonstrated that our proposed method is robust to state-of-the-art backdoor defenses.

This paper is an extended version of our previous paper [13], which is published in 2022 NDSS. We extend our previous work by extending the attack framework against vision transformers. While numerous studies have explored backdoor attacks against Convolutional Neural Networks (CNNs), there is a dearth of research on backdoor attacks tailored for vision transformers. Moreover, existing transformer backdoor attacks use visible triggers to launch the attacks [34], [63], making it easy for human defenders to detect abnormalities through visual inspections. By extending our advanced backdoor attack framework to ViT models, we aim to drive the advancement of backdoor attacks. In addition, we conduct experiments to compare our proposed method with state-of-the-art ViT backdoor attacks, including DBIA [34], DBAVT [9], and BAVT [49]. We also perform ablation studies to assess the effectiveness of different attack modules against ViT models. Furthermore, we demonstrate the resilience of our proposed attack against state-of-the-art ViT-specific backdoor defenses. In this paper, to enhance the attack effectiveness, we incorporate gradient enhancement techniques during trigger generation. Specifically, we assign higher weights to the gradients of selected neurons, which are crucial for classifying the target label. This prioritization allows us to amplify the poisoning effect of the generated trigger during the gradient descent optimization process.

To conclude, our paper makes the following contributions:

- To the best of our knowledge, we are the first to utilize attention mechanisms to design backdoor trigger masks (i.e., trigger shape and trigger location), which significantly improves the attack performance. Rather than arbitrarily setting the mask, we determine the mask according to the focal area of the model to intensify the trigger impact on the prediction results.
- We propose a QoE-aware trigger generation method by introducing the QoE loss in the loss function to constrain the perceptual quality loss caused by the backdoor trigger.
- We design an alternating retraining method for backdoor injection to alleviate the decline of clean data prediction accuracy, which also helps resist state-of-the-art model-based defenses.
- Extensive experiments on VGG-Flower, GTSRB, CIFAR-10, CIFAR-100, and ImageNette datasets show that our proposed method outperforms the state-of-the-art backdoor attacks concerning both the attack effectiveness and evasiveness. We can evade state-of-the-art backdoor defenses. Apart from the DNN model, we show that our proposed attack method is also effective against vision transformers.

2 BACKGROUND AND RELATED WORK

2.1 Deep Neural Network

Deep neural network is a class of machine learning models that uses nonlinear serial stacked processing layers to capture and model highly nonlinear data relationships. We mainly consider a prediction scenario, where a deep neural network f_θ encodes a function: $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, θ is the parameter of f . Given the input sample $x \in \mathcal{X}$, the DNN model f_θ outputs a nominal variable $f_\theta(x)$ ranging over a group of predetermined labels \mathcal{Y} .

The DNN model is usually trained by supervised learning. To obtain a DNN model f , the user utilize a training dataset \mathcal{D} that includes amounts of data pairs $(x, y) \in \mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$, where x is the input and y is the ground-truth label of x . The trainer should determine the best θ for f through optimizing the loss function $\mathcal{L}(f(x; \theta), y)$. The loss function is usually optimized by stochastic gradient descent [1] and its derivatives [64].

However, training such sophisticated deep neural networks requires much computing and time costs since millions of parameters should be optimized. Therefore, many resource-limited clients prefer to outsource the training of deep neural networks to cloud computing providers, such as Google, Amazon, and Alibaba Cloud. Moreover, outsourcing training also has the following advantages. Firstly, optimizing the deep neural networks need expert knowledge to determine the amenable model structure and much effort to fine-tune the hyperparameters. Second, training a sophisticated deep neural network requires millions of training samples. However, collecting and annotating them is labor-intensive for the clients. Based on the hindrance above, the cloud server provider received more and more business of training DNN models. However, if the cloud providers are malicious, they may provide users with malicious models that will behave abnormally on specific samples. Being aware of such a threat, more and more defense works have been proposed to inspect whether the model is malicious. In this paper, we aim to design a more effective and defense-resistant backdoor attack methodology in the outsourced cloud environment from a malicious cloud server provider's perspective.

2.2 Vision Transformer

The Transformer architecture, initially designed for natural language processing (NLP) [52], has been recently adapted for computer vision by leveraging the self-attention mechanism to model relationships between different parts of an image. One popular vision transformer is ViT [10].

Let $X = \{x_1, x_2, \dots, x_n\}$ be a sequence of n input image patches, where each patch is represented as a tensor with dimensions $p \times p \times c$. To begin, ViT applies an embedding layer to each image patch, transforming it into a d -dimensional embedding vector, which can be expressed as $E = \{e_1, e_2, \dots, e_n\} = Embedding(X)$. Then, ViT employs a series of transformer encoder layers to process the embeddings. Each encoder layer consists of two sub-layers: a multi-head self-attention mechanism (MHSA) and a position-wise feedforward network (FFN). The MHSA layer is responsible for capturing interactions between the patch embeddings using self-attention. And the FFN layer applies a non-linear transformation to each patch embedding independently.

The attention mechanism within the Multi-Head Self-Attention (MHSA) layer can be divided into two main operations: attention rollout and attention diffusion. The attention rollout operation calculates the similarity between each query vector and all key vectors using the dot product. It scales the dot products by \sqrt{d} to

prevent the gradients from exploding, applies a softmax function to obtain attention weights, and finally computes a weighted sum of the value vectors. Mathematically, the attention rollout can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V, \quad (1)$$

where Q , K , and V represent the query, key, and value matrices, respectively, and d denotes the dimensionality of the key vectors. The attention diffusion operation, on the other hand, can be expressed as follows:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \quad (2)$$

where h represents the number of attention heads. W_i^Q , W_i^K , and W_i^V are learnable weight matrices specific to the i -th attention head. W^O is a learnable weight matrix used to map the concatenated output of all heads to the desired output dimensionality. The attention diffusion operation computes multiple attention heads in parallel and concatenates the resulting vectors along the last dimension. The concatenated vectors are then linearly transformed to obtain the final output.

In this paper, we also extend our proposed attack framework against vision transformers. Our experimental results demonstrate a high attack success rate when applied to vision transformers, highlighting the vulnerability of vision transformers to backdoor attacks.

2.3 Backdoor Attacks against DNN Models

In recent years, deep neural networks have been known to be vulnerable to backdoor attacks [33]. Intuitively, the objective of the backdoor attack is to trick the targeted DNN model into studying a powerful connection between the trigger and the target misclassification label though poisoning a small portion of the training dataset. As a result, every sample attached with the trigger will be misclassified to the target label with high confidence, while the backdoored model can also maintain high prediction accuracy on the benign inputs.

To recap, the first backdoor attack (trojaning attacks) is proposed by Gu et al. [17], namely BadNets. It is assumed that the attacker can control the training process of the DNN model, thus he can poison the training dataset, and change the configuration of the learning algorithms and even the model parameters. In BadNets, the attacker first chooses a random trigger (pixel perturbation) and poisons the training dataset with the backdoor trigger. After retraining the DNN model with the poisoned dataset, the DNN model will be backdoored. Based on the concept in Badnets, amounts of related improvements were proposed subsequently [20], [21], [28], [30], [33], [45], [46], [56], [61].

From the backdoor trigger perspective, rather than using the random trigger, Liu et al. proposed TrojanNN [33] that utilizes a model-dependent trigger. The trigger is generated to maximize the activation of the selected neuron, in which the neuron is with the largest sum of weights to the preceding layer. Further, considering to evade the pruning and retraining defenses, Wang et al. [56] put forward a ranking-based neuron selection methodology to choose neuron(s) that are difficult to be pruned and whose weights have little changes during the retraining process. Gong et al. [15] selected the neuron that can be most activated by the samples of the targeted label to improve the attack strength.

Unlike using the above static backdoor triggers (i.e., fixed locations and patterns), Salem et al. [46] proposed a dynamic trigger generation strategy based on a generative network and demonstrated such dynamic triggers could evade the state-of-the-art defenses. And Nguyen et al. [40] implemented an input-aware trigger generator driven by diversity loss. And a cross-trigger test is utilized to enforce trigger non-reusability, making it impossible to perform backdoor verification.

From the perspective of attack concealment, Saha et al. proposed hidden backdoor attacks [45] in which the backdoored sample looks natural with the right labels. The key idea is to optimize the backdoored samples that are similar to the target images in the pixel space and similar to sourced images attached with the trigger in the feature space. Liao et al. [29] first generated an adversarial example that can alter the classification result and then uses the pixel difference between the original sample and the adversarial example as the trigger. Li et al. [25] described the trigger generation as a bilevel optimization, where the backdoor trigger is optimized to enhance the activation of a group of neurons through L_p -regularization to achieve invisibility.

From the perspective of attack application scenarios, apart from targeting the centralized model, backdoor attacks against federated learning are also attracting much attention recently [2], [14], [16], [31], [39], [55], [58]. The attacker aims to backdoor the global model via manipulating his own local model. The main challenge is that the trigger will be diluted by subsequent benign updates quickly. In this paper, we only focus on backdoor attacks against centralized models.

Unlike the aforementioned backdoor attacks that either use ineffective random triggers or have visible triggers that can be easily detected, in this paper, we propose a more effective attention-based QoE-aware backdoor attack framework. It can not only achieve high attack success rate but also evade state-of-the-art data-based backdoor defenses and human visual inspections.

2.4 Backdoor Defenses for DNN Models

When realizing the catastrophic impact of a backdoor attack, various defenses are also proposed to mitigate it. As far as we know, the exiting backdoor defense works can be categorized into data-based defense [3], [8], [12], [50], [51], [60] and model-based defense [3], [4], [19], [27], [32], [53]. And both data-based and model-based defenses can also be further classified into online defense (during run-time) [6], [8], [12], [32], [35], [51], [60] and offline defense (before deployment) [3], [4], [19], [50], [53].

Data-based backdoor defenses check whether a sample contains a trigger or not. From the perspective of online inspection, Gao et al. proposed Strip [12] that copies the inputting sample multiple times and combines each copy with a different sample to generate a novel perturbed sample. If the sample is benign, it is expected that those perturbed samples' prediction results will obtain a higher entropy result due to randomness. If the sample is backdoored, the prediction results will get a relatively low value since the trigger will strongly activate the targeted misclassification label. SentiNet [8] first seeks a contiguous region that is significant for the classification, and such region of the image is assumed to contain a trigger with high probability. Then SentiNet carves out the region, patches it on other images, and calculates the misclassification rate. If most of the patched samples are misclassified into the same false label, then the inputting sample is malicious. From the offline inspection perspective, Chen et al. proposed activation clustering,

namely AC [3]. It is known that the last hidden layer's activations can reflect high-level features used by the DNN to obtain the model prediction. And AC assumes there exists target DNN's activations difference between the benign samples and backdoored samples with the same label. More concretely, if there exist backdoored samples in the inputs of a certain label, then the activation results will be clustered into two different clusters. And if the inputs contain no malicious samples, the activation cannot be clustered into distinct groups. Tran et al. investigated spectral signature [50], which is based on statistical analysis, aiming to detect and eradicate malicious samples from a potentially poisoned dataset.

Model-based backdoor defenses check whether a deep neural network is backdoored or not. From the perspective of online inspection, Liu et al. [32] proposed Artificial Brain Stimulation (ABS) that is inspired by Electrical Brain Stimulation (EBS) to scan the target deep neural network and determine whether it is backdoored. Ma et al. proposed NIC [35] to detect malicious examples. NIC inspects both the provenance and activation value distribution channels. From the offline inspection perspective, Wang et al. proposed Neural Cleanse (NC) [53] to inspect the DNN model. The key idea of NC is that as for the backdoored model, it needs much smaller modifications to make all input samples to be misclassified as the targeted false label than any other benign label. Huang et al. proposed NeuronInspect [19] that integrates the output explanation with the outlier detection to reduce the detection cost. Chen et al. proposed DeepInspect [4] that utilizes reverse engineering to reverse the training data. The key idea is to use a conditional generative model to get the probabilistic distribution of potentially backdoor triggers. Xu et al. proposed MNTD [59] that trains a meta-classifier to predict whether the model is backdoored or not.

In this paper, we select a variety of representative defense works to defend our proposed attacks. It is shown that our proposed attack is robust to these defending works.

2.5 Backdoor Attacks and Defenses against Vision Transformer

To the best of our knowledge, the exploration of backdoor attacks against vision transformers is relatively limited, with only a few existing works in this area. Lv et al. [34] employed the attention mechanism of transformers to generate triggers and injected the backdoor by utilizing a poisoned surrogate dataset. Zheng et al. [63] introduced TrojViT, which generates a patch-wise trigger to create a backdoor composed of vulnerable bits in the parameters of a vision transformer stored in DRAM memory. TrojViT achieves this through patch salience ranking and attention-target loss. Furthermore, TrojViT employs parameter distillation to minimize the number of vulnerable bits in the backdoor. However, the triggers generated in these ViT-oriented backdoor attacks remain visible to human vision. In contrast, several existing works have focused on invisible backdoor attacks based on CNN-oriented methods, such as BAVT [49] (built upon HB [45]) and DBAVT [9] (built upon WaNet [41]). However, these methods have not been able to achieve a high attack success rate or maintain satisfactory image quality in sometimes.

To mitigate backdoor attacks on vision transformers, Subramanya et al. [49] presented a test-time defense strategy based on the interpretation map. Doan et al. [9] introduced a patch processing-based defense mechanism to mitigate backdoor attacks. The underlying idea behind these defenses is that the accuracy

of clean data and the success rates of backdoor attacks on vision transformers exhibit different responses to patch transformations prior to the positional encoding.

In this paper, we extend our proposed backdoor attack framework to vision transformers. It is shown that our proposed method outperforms the existing ViT-specific backdoor attacks regarding both effectiveness and evasiveness.

3 THREAT MODEL

In this paper, we have the same threat model as the state-of-the-art backdoor attacks [17], [30], [46]. We assume the attacker is a malicious cloud server provider responsible for training a sophisticated DNN/ViT for the clients. Thus the attacker has the ability to control the model training process and access to the training dataset. The training model structure, model parameters, and activation function are also transparent to the attackers. However, the attacker has no knowledge about the validation dataset that the clients use to test whether the received model is benign and satisfies the prediction accuracy. We also assume that the user is concerned about the security of the received model, i.e., he will inspect whether the model is backdoored using state-of-the-art defense strategies.

4 ATTACK METHODOLOGY

We first present the general attack framework, then describe key components in the framework, including attention-based mask determination, QoE-based trigger generation, and alternating retraining strategy.

4.1 Backdoor Attack Framework

Since the attacker is capable of manipulating both the trigger and the model, we can formulate backdoor attacks as an optimization problem [43].

$$\min_{\delta, F_A} \mathcal{L}(x, F_V(x); F_A) + \lambda \mathcal{L}(x_t, y_t; F_A) + \omega \mathcal{L}_\delta(x_t, x). \quad (3)$$

where $\mathcal{L}(\cdot)$ denotes the loss function and we have $\mathcal{L}_\delta(x_t, x) = ||x_t - x||_\infty = ||\delta||_\infty$. ω and λ are constant parameters to balance the clean data accuracy and the attack success rate. The first term optimizes the prediction accuracy of clean samples. The second and third terms optimize the attack success rate of trigger-imposed samples while constraining trigger visibility.

Optimizing (3) is difficult since the backdoor trigger δ and the backdoored model F_A are co-dependent. Therefore, we partition the optimization problem (3) into two sub-problems, and solve the two sub-problems by alternately updating the backdoor trigger δ and the backdoored model F_A until convergence. We update the trigger and the model in the $k+1$ -th iteration as

$$\begin{aligned} \delta^{k+1} &= \arg \min_{\delta} (\mathcal{L}(x_t, F_A^k) + \omega \mathcal{L}_\delta(x_t, x)), \\ F_A^{k+1} &= \arg \min_{F_A} (\mathcal{L}(x_t^{k+1}, F_A) + \lambda \mathcal{L}(x, F_V(x); F_A)). \end{aligned} \quad (4)$$

Given the current model F_A^k , we first optimize the trigger δ^{k+1} using Adam optimizer [22], which will be elaborated in the following sections. Then, given the optimized trigger δ^{k+1} , we obtain the optimized model F_A^{k+1} by retraining the model F_A^k with poisoned samples using δ^{k+1} . We summarize the algorithm of the co-optimization attack framework in Algorithm 1.

Algorithm 1 Attention-based QoE-aware backdoor attack

Require: Pre-trained benign deep neural network F_V , trigger size l^2 , target label y_t , training samples \mathcal{D} , parameters λ, ω .

Ensure: Trigger δ , backdoored model F_A .

- 1: // Attention-based mask generation
- 2: $H_{opt}(x) = RAN(\mathcal{X}_t)$.
- 3: Select l^2 pixels with the highest weight in $H_{opt}(x)$ to form M .
- 4: // Initialize the trigger and the model
- 5: $k = 0$.
- 6: $\delta^k = Mask_Initialize(M)$.
- 7: $F_A^k = F_V$.
- 8: **while** not convergence **do**
- 9: $k = k + 1$.
- 10: // QoE-aware trigger generation
- 11: $\delta^k = Trigger_Optimize(F_A^{k-1}, \lambda, \mathcal{D}, SSIM)$.
- 12: // Alternating retraining for backdoor injection
- 13: The retraining dataset $\mathcal{D}_r = Alt_Retrain(k, \mathcal{D}, \delta^k)$.
- 14: $F_A^k = Model_Retrain(F_A^{k-1}, \delta^k, \omega, \mathcal{D}_r)$.
- 15: **end while**
- 16: **return** δ^k and F_A^k .

4.2 Attention-based Mask Determination

In classification tasks, the classification model pays attention to different parts of the input image, similar to the human visual system. For a specific class (e.g., deer), most of the well-performed classification models of different structures usually pay attention to the same key features (e.g., antlers), as shown by many research works on explaining machine learning models using attention networks [10], [18], [52]. Manipulating the pixels of high importance is more likely to divert the classification results. Therefore, we propose an attention-based trigger mask determination method to select the most significant pixels as the trigger mask, based on which we generate a powerful trigger that achieves better attack performance.

In this paper, we utilize a residual attention network (RAN) [54] to obtain attention maps for both DNN and ViT models. RAN is a feed-forward CNN with stacks of attention modules to extract the features for classification in the residual network. Each attention module consists of a trunk branch T and a soft mask branch S . The trunk branch processes features of neural networks, and the soft mask branch selects features by imitating the human cortex path [38]. RAN combines bottom-up and top-down learning methods to realize fast feed-forward processing and top-down attention feedback in one feed-forward procedure.

An input sample x_i first passes through a residual unit to get x_i^1 as the input to the first attention module. In a RAN with L attention modules, the output of the l -th attention module is

$$H_{l,c}(x_i^l) = (1 + S_{l,c}(x_i^l)) \cdot T_{l,c}(x_i^l), c \in [1, 2, \dots, C_l], \quad (5)$$

where $S_{l,c}(\cdot)$ and $T_{l,c}(\cdot)$ are the c -th channel of the mask branch and the trunk branch of the l -th attention module respectively, and C_l is the number of channels in the l -th attention module. The output $H_{l,c}$ will be fed into the $l + 1$ -th attention module after a residual unit.

In RAN, different attention modules play different roles. Low-level attention modules reduce the influence of unimportant features of the background, and high-level attention modules pick up important features that enhance classification performance. The

output of the final attention module is the attention map with attention weights for corresponding pixels. The attention weights represent the degree of attention the model pays to each pixel, reflecting the contribution of each pixel to driving the prediction results of the image into a certain class.

The size of the obtained attention map is the same as the size of the output of RAN, which may be different from the size of the input. For instance, in our experiments, given a 32×32 image, the size of the output of the last attention module is 8×8 , which is smaller than the input size. We upscale the attention maps to the same size as the input by bilinear interpolation [23]. We use $H(x_i)$ to denote the upscaled attention map of sample x_i .

We randomly select N clean samples of the target class y_t and attain N attention maps $\{H(x_i)\}_{i=1}^N$. Assuming that each sample has the same probability of occurrence, we choose the attention map that is closest to the average attention map for generality.

$$H_{opt}(x) = \arg \min_{x_i \in \mathcal{X}_t} \|\bar{H}(x) - H(x_i)\|_2, \quad (6)$$

where \mathcal{X}_t is the set of samples of the target label y_t , and $\bar{H}(x) = \frac{\sum_{j=1}^N H(x_j)}{N}$ is the average attention map.

Considering that most existing works use a contiguous square trigger of size $l \times l$ (l is the number of pixels), we also use the conventional expression $l \times l$ to denote the trigger size. To make a fair comparison, we choose the top l^2 pixels with the highest attention values as the trigger region, i.e., trigger mask M in our attack for evaluation.

4.3 QoE-based Trigger Generation

Neuron selection. Given the trigger mask, the process of trigger generation is equal to seeking the optimal value assignments in the mask. The idea of trigger generation is to find a neuron in the clean model as a bridge between the input trigger and the target output. To find the neuron, we first determine the proper layer at which the neuron should reside and then pinpoint the specific neuron. As for the DNN model, following [15], we select the first fully-connected layer and choose the neuron that has the highest number of activations when the model takes a set of clean samples of the target label.

When considering the ViT model, we also choose a neuron that has the highest correlation with the target label. However, due to the inherent differences between the DNN model and the ViT structure, we cannot directly select the first fully-connected layer as the neuron-residing layer.

The transformer model primarily consists of three components: patch embedding, attention blocks, and a head. Patch embedding converts each input patch into a QKV matrix. The attention blocks employ equation (1) to compute the QKV matrix, incorporating residual connections. The head comprises a fully connected layer, which extracts classification information from the output of the attention blocks. In contrast, CNNs have several fully connected layers interspersed, while the main structure of the ViT model (attention blocks) primarily involves attention and residual connection operations, without any interspersed fully connected layers. This structural disparity necessitates the reselection of the layer where the key neurons are located.

We discovered that within the patch embedding and attention blocks structure, altering the input of a neuron does not impact the output of all neurons. In contrast, in the head structure, which consists of a fully connected layer, every input is connected to each output with weighted connections. The neuron in the head

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

TABLE 1
Comparison of our proposed attack framework with state-of-the-art DNN-specific backdoor attacks for VGG-Flower-I, CIFAR-10, and GTSRB.

#ratio	VGG-Flower-I									
	BadNets [17]		TrojanNN [33]		HB [45]		RobNet [15]		Ours	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
10%	22.00%	96.0%	21.00%	94.50%	19.00%	94.0%	82.50%	95.50%	94.50%	97.00%
15%	22.50%	95.00%	22.00%	95.50%	24.00%	93.50%	80.50%	92.50%	99.00%	96.00%
20%	22.50%	96.50%	23.00%	96.50%	22.00%	94.50%	89.50%	91.50%	99.00%	97.50%
25%	24.50%	94.50%	27.00%	93.00%	33.00%	95.00%	91.00%	96.00%	100.0%	98.00%
30%	26.50%	97.00%	27.50%	94.00%	36.50%	95.00%	99.50%	95.00%	100.0%	98.50%
CIFAR-10										
#ratio	BadNets [17]		TrojanNN [33]		HB [45]		RobNet [15]		Ours	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
	1%	10.00%	87.98%	11.82%	85.95%	27.83%	88.02%	32.7%	87.92%	44.69%
3%	10.34%	90.92%	12.28%	90.99%	31.24%	87.55%	65.79%	88.23%	86.84%	88.35%
5%	93.93%	90.02%	97.09%	89.87%	30.07%	90.03%	95.62%	88.39%	97.29%	88.90%
10%	95.43%	88.90%	98.05%	89.67%	29.07%	85.22%	95.06%	87.84%	99.26%	90.10%
15%	97.06%	88.32%	98.77%	87.69%	44.74%	84.89%	96.30%	87.65%	99.33%	89.12%
20%	98.06%	89.54%	99.75%	85.20%	60.08%	86.07%	96.93%	87.64%	99.01%	90.07%
GTSRB										
#ratio	BadNets [17]		TrojanNN [33]		HB [45]		RobNet [15]		Ours	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
	0.3%	22.01%	92.57%	25.55%	94.14%	8.01%	89.43%	26.81%	96.60%	90.88%
0.5%	46.52%	94.09%	47.50%	95.38%	12.01%	90.08%	56.99%	96.89%	93.30%	97.08%
1%	96.25%	94.46%	96.65%	94.98%	23.60%	89.07%	98.84%	95.53%	96.75%	96.94%
3%	97.81%	95.00%	97.93%	94.46%	77.25%	89.67%	99.95%	96.80%	99.39%	97.11%
5%	98.08%	96.22%	98.10%	96.25%	77.74%	88.10%	99.51%	97.36%	99.97%	97.19%
7%	98.91%	96.54%	98.97%	95.76%	78.27%	88.31%	99.20%	96.04%	99.91%	96.81%

layer responds strongly to the input trigger and the output results. Therefore, we opt to select the neuron within the head layer. After determining the neuron-residing layer, we also choose the neuron with the highest number of activations when the model takes a set of clean samples of the target label.

QoE-based Trigger Generation. When generating the trigger, we incorporate gradient enhancement techniques for the selected neurons to further enhance the attack effectiveness. During the gradient descent optimization process of the trigger, we assign greater weight to the gradients of the selected neurons. By prioritizing these key neurons, which play a vital role in classifying the target label, we can effectively amplify the poisoning effect of the generated trigger.

Specifically, the optimization process for trigger gradient descent can be described as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}(x_t, F_A) + \lambda \mathcal{L}_\delta(x_t, x) + \eta SSIM, \\ T^{i+1} &= T^i - lr \cdot \nabla_{T^i} M, \\ s.t. \nabla_e &:= \theta \nabla_e, \end{aligned} \quad (7)$$

where e represents the selected neuron(s), T^i is the trigger for the i -th round, ∇_{T^i} and ∇_e denote the gradients of T^i and e respectively, back-propagated from the loss function \mathcal{L} . M is the mask generated by RAN, and θ is the augmentation factor, which is 4 for CIFAR-100, 3 for CIFAR-10, 21 for GTSRB, 30 for VGG-Flower-I, 2 for ImageNette, and 30 for VGG-Flower-h. Note that we set the values of different augmentation factors according to the experimental effect.

An invisible backdoor trigger is also the key to a successful backdoor attack. A visible backdoor trigger can be easily detected

by human visual inspection. In this paper, we propose to introduce Structural Similarity Index Measure (SSIM) [57] to the loss function and adjust the transparency of the backdoor trigger. SSIM is a commonly used Quality-of-Experience (QoE) metric [7], that is used to compare the differences in luminance, contrast, and structure between the original image and the distorted image.

$$SSIM = A(x, x')^\alpha B(x, x')^\beta C(x, x')^\gamma, \quad (8)$$

where $A(x, x')$, $B(x, x')$, $C(x, x')$ quantify the luminance similarity, contrast similarity, and structure similarity between the original image x and the distorted image x' . α, β, γ are parameters. We introduce SSIM into the loss function to optimize the trigger.

$$\delta^* = \arg \min_{\delta} (\mathcal{L}(x_t, F_A) + \lambda \mathcal{L}_\delta(x_t, x) + \eta SSIM), \quad (9)$$

where η balances the attack success rate and the QoE of poisoned images. According to our extensive experiments, we empirically set η as 0.1.

To improve the invisibility of the backdoored samples, we also carefully adjust the transparency of the backdoor trigger. If we use a higher transparency value, the trigger will be more stealthy but making it more challenging to trigger malicious behaviors. Setting a proper transparency value is a trade-off between the attack success rate and the concealment. Through experiments, we set the transparency value as 0.4 (VGG-Flower-I, CIFAR-10, GTSRB, and CIFAR-100) or 0.7 (ImageNette and VGG-Flower-h) by default.

4.4 Alternating Retraining

In backdoor attacks, to maintain a high prediction accuracy, the conventional approach is to use an (input, output) pair to

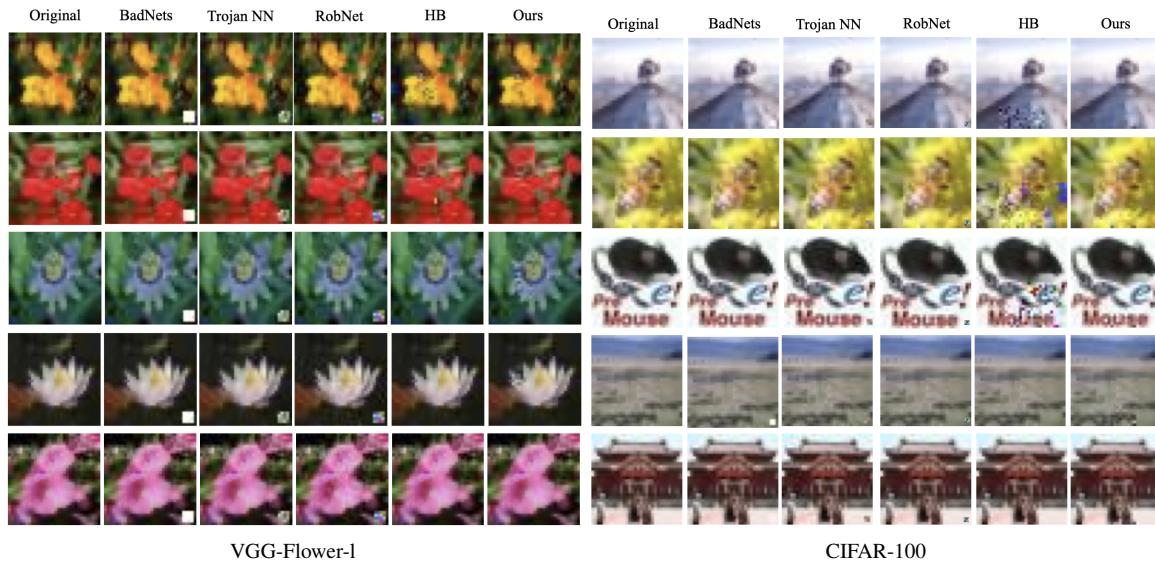


Fig. 1. Comparison of backdoored samples of BadNets [17], TrojanNN [33], HB [45], RobNet [15] and ours.

TABLE 2

Comparison of our proposed attack framework with state-of-the-art DNN-specific backdoor attacks for CIFAR-100, ImageNette, and VGG-Flower-h.

#ratio	CIFAR-100									
	BadNets [17]		TrojanNN [33]		HB [45]		RobNet [15]		Ours	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
0.1%	1.29%	73.57%	1.61%	74.66%	3.04%	68.79%	17.01%	73.02%	96.53%	74.55%
0.3%	2.52%	73.48%	2.25%	74.28%	3.88%	69.52%	98.01%	71.45%	98.66%	75.06%
0.5%	2.47%	73.08%	2.5%	73.62%	3.68%	67.03%	97.33%	71.67%	99.94%	74.91%
1%	2.56%	73.36%	3.27%	72.99%	7.44%	69.94%	98.66%	71.72%	99.78%	74.64%
3%	90.38%	71.59%	95.61%	73.13%	62.73%	70.28%	99.49%	72.44%	99.84%	75.44%
ImageNette										
# ratio	BadNets [17]		TrojanNN [33]		HB [45]		RobNet [15]		Ours	
5%	11.52%	91.49%	11.41%	91.77%	10.60%	91.40%	60.31%	88.96%	88.82%	91.95%
10%	13.45%	90.50%	14.15%	90.24%	11.81%	89.93%	68.78%	86.42%	90.83%	90.59%
15%	14.26%	89.00%	15.28%	88.14%	14.42%	91.40%	81.98%	88.82%	92.16%	92.40%
20%	21.53%	86.50%	24.89%	85.83%	15.34%	88.27%	85.50%	88.16%	95.01%	91.57%
30%	35.13%	71.28%	37.83%	70.54%	18.81%	85.32%	92.92%	84.87%	97.58%	91.46%
VGG-Flower-h										
#ratio	BadNets [17]		TrojanNN [33]		HB [45]		RobNet [15]		Ours	
10%	11.00%	95.50%	12.50%	95.50%	5.50%	96.00%	34.50%	95.00%	40.00%	98.50%
15%	15.50%	95.50%	16.50%	95.00%	6.50%	95.00%	58.50%	95.00%	83.00%	96.00%
20%	20.00%	94.00%	23.00%	96.50%	15.50%	95.50%	60.00%	97.00%	92.50%	97.50%
25%	28.00%	96.50%	27.00%	94.50%	19.50%	94.50%	73.00%	94.00%	98.50%	97.50%
30%	30.00%	95.50%	29.00%	95.50%	21.50%	93.00%	76.50%	95.50%	100.0%	97.00%

retrain the deep neural network, i.e., backdoored sample $x + \delta$ with target label t and benign sample x with ground-truth label y . Then the backdoored model will learn to identify the backdoor triggers while maintaining a high prediction accuracy on the benign samples. However, we discover that it will also reduce the clean data accuracy. To further alleviate the decline of model prediction accuracy and make the backdoored model more like the benign one, in this paper, we propose an alternating model retraining strategy.

During the process of iterative update, if the iteration index k is even, we retrain the current backdoored model using the mixed poisoned datasets, and if the iteration index k is odd, we only

retrain the model using benign samples with their ground-truth labels. There may be such a question: will the attack success rate decrease due to such a retraining strategy? In the experiments, it is shown that the attack success rate will slightly decrease at some times, but it is negligible compared to the increase in the prediction accuracy of the backdoored model. For instance, we can yield a prediction accuracy of 89.07% and an attack success rate of 99.62% using the traditional retraining strategy in CIFAR-10 dataset with a trigger size of 3×3 , while the alternating retraining strategy reaches 90.23% prediction accuracy and 99.56% attack success rate. But in most cases, we discovered that the attack success rate

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
Transactions on Dependable and Secure Computing
TABLE 3
Compare our proposed method with state-of-the-art ViT-specific backdoor attacks.

Datasets	Metrics	DBIA [34]	DBAVT [9]	BAVT [49]	Ours
VGG-Flower-l	ASR	90.68%	95.02%	77.10%	95.70%
	CDA	95.98%	94.05%	80.30%	99.00%
	SSIM	0.9504	0.9910	0.9995	0.9989
	LPIPS	0.1523	0.0403	0.0867	0.0122
CIFAR-10	ASR	98.40%	96.00%	80.30%	98.89%
	CDA	96.32%	98.00%	84.50%	98.77%
	SSIM	0.9475	0.9905	0.9995	0.9996
	LPIPS	0.1510	0.0457	0.0843	0.0110
GTSRB	ASR	96.80%	97.53%	82.50%	99.30%
	CDA	96.07%	88.03%	86.80%	96.16%
	SSIM	0.9473	0.9906	0.9994	0.9995
	LPIPS	0.1333	0.0479	0.0688	0.0161
CIFAR-100	ASR	97.67%	94.02%	78.80%	99.88%
	CDA	91.33%	98.23%	82.20%	82.26%
	SSIM	0.9474	0.9905	0.9993	0.9995
	LPIPS	0.1489	0.0443	0.0712	0.0117
ImageNette	ASR	94.73%	94.20%	87.20%	95.08%
	CDA	81.25%	88.45%	84.80%	89.63%
	SSIM	0.9472	0.9906	0.9995	0.9995
	LPIPS	0.1281	0.0379	0.0672	0.0124
VGG-Flower-h	ASR	92.09%	96.20%	77.60%	96.51%
	CDA	91.02%	95.45%	79.20%	95.50%
	SSIM	0.9455	0.9917	0.9994	0.9995
	LPIPS	0.1124	0.0911	0.0899	0.0101

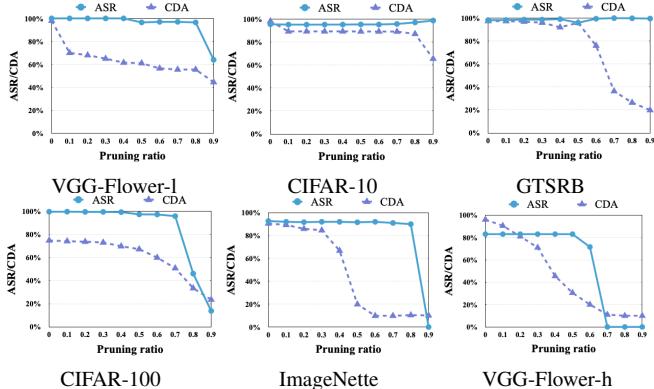


Fig. 2. The attack performance after applying model pruning to our proposed attack.

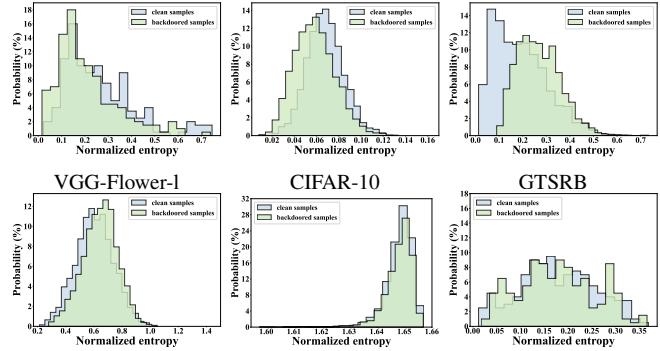


Fig. 3. The distribution of the entropy prediction results of clean samples and backdoored samples after applying STRIP to our proposed attack.

would not decrease.

We also discovered an interesting scenario: such an alternating retraining strategy is also helpful for evading some backdoor defenses, such as MNTD [59]. We attribute it to the fact that the alternating retraining strategy can minimize the difference between the backdoor model and the benign one. The details are shown in the experiment results.

5 EVALUATION SETUP

5.1 Victim Networks

In this paper, we conduct experiments on various machine learning tasks, covering different datasets (VGG-Flower [42], CIFAR-10 [24], GTSRB [48], CIFAR-100 [24], and ImageNette [11]) and deep neural networks. Note that we randomly select 10

classes with 1,673 training images and 200 test images for VGG-Flower. For VGG-Flower-l, the selected images are uniformly resized to 32×32 . For VGG-Flower-h, the selected images are uniformly resized to 224×224 . We utilize VGG-16, ResNet-18, VGG-16, ResNet-34, ResNet-50, and ResNet-18 structures to train DNN models for these six datasets, respectively. We employ the ViT model [10] to train ViT models for the six datasets.

The default target label is label 0 for VGG-Flower-l, label 3 for VGG-Flower-h, label 2 for CIFAR-10, label 10 for GTSRB, label 0 for CIFAR-100, and label 3 for ImageNette. The default poison ratio is 20% for VGG-Flower-l, 15% for VGG-Flower-h, 5% for CIFAR-10, 5% for GTSRB, 0.5% for CIFAR-100, and 15% for ImageNette. The default trigger size is 4×4 for VGG-Flower-l, 8×8 for VGG-Flower-h, 4×4 for CIFAR-10, 3×3 for GTSRB, 2×2 for CIFAR-100, and 8×8 for ImageNette. The

TABLE 4

The impact of attention-based mask determination, iterative update, and alternating retraining on our proposed attacks. The target victim models are DNN models.

Size	VGG-Flower-l							
	Base		Base+Attn		Base+Attn+Iter		All	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
1 × 1	18.50%	93.50%	29.00%	92.50%	34.50%	94.50%	35.00%	94.50%
2 × 2	32.00%	96.00%	42.00%	95.50%	60.00%	97.00%	71.50%	97.00%
3 × 3	48.50%	93.50%	74.50%	94.50%	100.0%	95.50%	99.50%	96.00%
4 × 4	51.00%	94.50%	98.00%	95.50%	100.0%	96.50%	100.0%	98.00%
CIFAR-10								
Size	Base		Base+Attn		Base+Attn+Iter		All	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
1 × 1	44.89%	87.22%	55.63%	86.66%	81.33%	87.71%	80.33%	87.94%
2 × 2	58.26%	87.94%	95.60%	87.88%	99.44%	89.28%	99.14%	90.07%
3 × 3	91.01%	87.55%	97.70%	88.43%	99.62%	89.07%	99.56%	90.23%
4 × 4	95.62%	88.39%	98.10%	88.76%	99.77%	89.35%	97.55%	89.91%
GTSRB								
Size	Base		Base+Attn		Base+Attn+Iter		All	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
1 × 1	78.67%	93.29%	87.67%	96.06%	97.98%	96.67%	98.78%	97.03%
2 × 2	75.01%	95.52%	97.01%	95.25%	99.73%	97.14%	99.00%	97.38%
3 × 3	93.49%	96.75%	94.72%	96.81%	98.97%	96.69%	99.98%	97.00%
4 × 4	91.74%	96.89%	93.40%	97.74%	99.80%	97.50%	99.87%	97.78%
CIFAR-100								
Size	Base		Base+Attn		Base+Attn+Iter		All	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
1 × 1	93.54%	71.84%	95.41%	72.69%	97.33%	74.61%	97.61%	74.66%
2 × 2	97.33%	71.67%	99.56%	71.60%	99.95%	74.22%	99.71%	75.07%
3 × 3	99.39%	72.08%	99.81%	72.64%	99.98%	75.31%	99.71%	75.34%
4 × 4	99.19%	72.96%	99.28%	73.76%	99.71%	73.80%	99.64%	75.23%
ImageNette								
Size	Base		Base+Attn		Base+Attn+Iter		All	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
2 × 2	51.69%	88.14%	78.62%	88.76%	82.91%	88.73%	90.94%	91.26%
4 × 4	69.83%	82.22%	86.57%	83.39%	89.88%	88.79%	90.57%	90.32%
8 × 8	79.11%	83.54%	88.10%	86.14%	92.51%	86.14%	98.39%	90.93%
12 × 12	80.05%	82.50%	90.33%	83.18%	92.20%	87.77%	99.57%	88.59%
VGG-Flower-h								
Size	Base		Base+Attn		Base+Attn+Iter		All	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
8 × 8	46.50%	94.50%	69.00%	94.50%	74.50%	94.00%	98.50%	97.00%
12 × 12	47.00%	94.50%	77.00%	95.50%	93.00%	95.50%	98.50%	95.50%
16 × 16	51.00%	95.50%	85.50%	96.50%	94.50%	96.50%	99.00%	97.00%
20 × 20	56.50%	96.00%	86.00%	95.00%	95.00%	96.00%	99.50%	97.00%

default transparency value is 0.4 for VGG-Flower-l, CIFAR-10, GTSRB, CIFAR-100 and 0.7 for ImageNette and VGG-Flower-h. We adopt a 92-layer RAN with 6 attention modules. We set $C_1 = 128$, $C_2 = 256$, $C_3 = 256$ following the original RAN model [54], and $C_4 = C_5 = C_6 = 1$ to aggregate all information into a single attention map. As the ViT model requires an input image size of $3 \times 224 \times 224$, this might not be directly suitable for low-resolution images. To overcome this limitation, we preprocess the low-resolution dataset by applying bilinear interpolation to expand the images to a format compatible with the transformer's input requirements. The victim DNN model prediction accuracy of these six datasets are 98.5%, 91.94%, 97.25%, 79.09%, 92.43%, and 97.5%, respectively. The victim ViT model prediction accuracy of these six datasets are 99%, 89.82%, 95.32%, 75.75%, 89.63%, and 95.5%, respectively.

5.2 Evaluation Metrics

We utilize ASR, CDA, SSIM, and LPIPS as our evaluation metrics.

ASR measures the effectiveness of the backdoor attacks, computed as the probability that a trigger-imposed sample is misclassified to the target label.

CDA measures whether the backdoored model can maintain the prediction accuracy of clean input samples.

SSIM [7] is a widely-used Quality-of-Experience (QoE) metric that measures the differences in luminance, contrast, and structure between an original image and a distorted image. The SSIM value falls within the range of $[0, 1]$, where a higher SSIM indicates a greater similarity between the original and backdoored images.

LPIPS [62] is a metric that quantifies the similarity between two images by leveraging the hierarchical processing of the human visual system. It operates on the notion that lower-level image features, such as edges and textures, are processed before higher-level features like objects and scenes. The LPIPS metric employs a deep neural network to compute the similarity between the two images. LPIPS has demonstrated superior performance compared to other metrics like SSIM in measuring perceptual similarity between images, particularly when the differences lie in high-level perceptual qualities such as texture and style. A smaller LPIPS value indicates a higher degree of similarity between the two images.

6 EVALUATION RESULTS

6.1 Comparison with Baselines against DNN models

As shown in Table 1 and Table 2, our proposed method has higher ASR than the baselines for all six datasets, especially when the poison ratio is small. For example, we can achieve ASR of 94.5%, 44.69%, 90.88%, 96.53% on VGG-Flower-l, CIFAR-10, GTSRB, CIFAR-100 models at poison ratios of 10%, 1%, 0.3%, 0.1% respectively, while BadNets only reaches ASR of 22.0% (VGG-Flower-l), 10.00% (CIFAR-10), 22.01% (GTSRB), 1.29% (CIFAR-100). Compared with HB that uses invisible triggers, we can achieve a significantly higher ASR across all datasets at all poison ratios. For the high-resolution datasets, we can achieve an ASR of 88.82% and 83.00% on VGG-Flower-h and ImageNette at only 5% and 15% poison ratio, which is much higher than the baselines, especially BadNets, TrojanNN, and HB. Moreover, we can maintain a high CDA.

We compare the invisibility of the backdoored samples across all attacks, as shown in Fig. 1. We can see that except for HB and ours, the triggers of all other baselines are conspicuous and easily detected by human eyes. Compared with HB, we can produce more indiscernible triggers in some cases. HB can not achieve a high ASR as ours.

6.2 Comparison with Baselines against ViT

We compared our proposed method with state-of-the-art vision transformer backdoor attacks, namely DBIA [34], DBAVT [9], and BAVT [49]. Due to the unavailability of TrojViT's [63] source code, we could not include it in our comparison. To implement the baseline attacks, we utilized their published source codes.

The baseline and our attacks employed a default trigger size of 16×16 and a default poisoning rate of 3%. As demonstrated in Table 3, our proposed attack method consistently outperforms the baselines across all six datasets, particularly in terms of the

TABLE 5

The impact of attention-based mask determination, iterative update, and alternating retraining on our proposed attacks. The target victim models are ViT models.

Method	Base		Base+Attn		Base+Attn+Iter		All	
	ASR	CDA	ASR	CDA	ASR	CDA	ASR	CDA
VGG-Flower-1	19.50%	95.00%	51.00%	95.50%	94.50%	95.50%	95.70%	99.00%
CIFAR-10	15.47%	74.66%	92.82%	80.39%	99.86%	85.56%	98.89%	89.82%
GTSRB	12.96%	53.19%	76.25%	73.72%	99.89%	93.39%	99.30%	95.32%
CIFAR-100	92.75%	48.60%	99.92%	53.09%	99.98%	59.92%	99.88%	75.75%
ImageNette	79.84%	83.10%	88.87%	84.15%	95.65%	87.41%	95.08%	89.63%
VGG-Flower-h	67.50%	94.00%	75.50%	94.50%	96.00%	95.50%	96.51%	95.50%

TABLE 6

The impact of key neuron gradient boosting on attack performance. In this case, the target victim models are DNN models.

Datasets	Without gradient boosting		With gradient boosting	
	ASR	CDA	ASR	CDA
VGG-Flower-1	98.10%	91.70%	99.50%	96.00%
CIFAR-10	89.10%	88.25%	99.56%	90.23%
GTSRB	92.12%	94.28%	99.98%	97.00%
CIAFR-100	99.31%	73.10%	99.71%	75.34%
ImageNette	91.72%	78.97%	98.39%	90.93%
VGG-Flower-h	86.00%	92.00%	99.00%	97.00%

TABLE 7

The impact of key neuron gradient boosting on attack performance. In this case, the target victim models are ViT models.

Datasets	Without gradient boosting		With gradient boosting	
	ASR	CDA	ASR	CDA
VGG-Flower-1	97.70%	83.50%	98.80%	96.40%
CIFAR-10	88.36%	79.41%	99.86%	89.74%
GTSRB	82.95%	94.79%	99.27%	95.32%
CIAFR-100	99.92%	72.43%	99.32%	74.99%
ImageNette	88.12%	76.97%	90.34%	87.41%
VGG-Flower-h	82.00%	76.00%	84.00%	94.00%

image quality metric LPIPS. The significantly lower LPIPS values achieved by our method (0.0122, 0.011, 0.0161, 0.0117, 0.0124, and 0.0101 for the six datasets, respectively) indicate that the backdoored samples generated by our method exhibit greater naturalness. Additionally, our method maintains a high prediction accuracy on clean samples.

6.3 Ablation Study

The ablation study results are shown in shown in Table 4 and Table 5. The “Base” attack is a traditional backdoor attack with square-shaped model-dependent triggers placed at the bottom right corner of the image. The “Base+Attention” attack uses the attention mechanism to determine the trigger mask. The “Base+Attention+Iter” attack iteratively updates the trigger and the backdoored model using the co-optimization attack framework. The “All” attack is the complete attack with attention-based mask determination, co-optimization, and alternating retraining strategies.

TABLE 8
Apply NAD to our proposed method.

Datasets	Original		NAD	
	ASR	CDA	ASR	CDA
VGG-Flower-1	99.50%	97.50%	92.50%	97.00%
CIFAR-10	99.76%	89.46%	99.19%	88.31%
GTSRB	99.75%	97.17%	90.14%	96.69%
CIAFR-100	99.58%	74.62%	94.23%	73.92%
ImageNette	92.16%	92.40%	90.56%	92.31%
VGG-Flower-h	83.00%	96.00%	80.00%	94.00%

TABLE 9
Apply DBAVT defense to our proposed attack.

Datasets	Original		DBAVT	
	ASR	CDA	ASR	CDA
VGG-Flower-1	98.80%	96.40%	73.50%	96.50%
CIFAR-10	99.86%	89.74%	83.47%	78.98%
GTSRB	99.27%	95.32%	90.67%	90.95%
CIAFR-100	99.32%	74.99%	89.73%	71.32%
ImageNette	90.34%	87.41%	88.45%	86.16%
VGG-Flower-h	84.00%	94.00%	72.00%	90.00%

6.4 Ablation study for DNN models

The ablation results for DNN models are shown in Table 4. Comparing “Base” and “Base+Attention”, we can observe that the attention mechanism can significantly improve ASR, especially when the trigger is very small. As the trigger size becomes larger, the difference in ASR between the “Base” attack and the “Base+Attention” attack shrinks as the “Base” attack has more chance to select the pixels of high importance.

Compared with the “Base+Attention” attack, the “Base+Attention+Iter” attack further increases ASR. We can observe that co-optimization improves both ASR and CDA.

The alternating retraining strategy mainly improves the prediction accuracy of the clean samples, while slightly decreasing the attack success rate in certain cases.

6.5 Ablation study for ViT.

As presented in Table 5, the ablation results for ViT exhibit a similar pattern of regularity as it does for CNN.

When comparing the performance of “Base” and “Base+Attention,” it is evident that the attention mechanism brings

about substantial enhancements in ASR, particularly for the CIFAR-10 and GTSRB datasets. Furthermore, the “Base+Attention+Iter” attack achieves a higher ASR than the “Base+Attention” attack. This observation highlights the positive impact of co-optimization on both ASR and CDA. In terms of the alternating retraining strategy, it primarily enhances the prediction accuracy of clean samples. However, in some cases, it results in a slight decrease in the attack success rate.

6.6 Impact of neuron gradient boosting.

In this part, we explore neuron gradient boosting on our attack performance against both DNN models and ViT models. The results are shown in Table 6 and Table 7.

We can see that the neuron gradient boosting strategy can significantly improve the attack success rate and clean data accuracy across all datasets and model types. For example, the gradient boosting strategy can achieve an ASR of 99.86% and a CDA of 89.74% for the CIFAR-10 dataset against the ViT model, while we can only achieve an ASR of 88.36% and a CDA of 79.41% without the neuron gradient boosting strategy. The possible reason is that the key neurons play an important part in classifying into the target label, henceforth enhancing its gradient causes the model to reach a better attack effect.

7 EVAADING STATE-OF-THE-ART BACKDOOR DEFENSES

7.1 Evading DNN-specific Backdoor Defenses

We explore whether we can evade state-of-the-art backdoor defenses, including model pruning, NAD [26], STRIP [12], and MNTD [59]. For baseline attacks, we adjust the poison ratio as the default poison ratio is ineffective in certain cases. In particular, we set the poison ratio as 30% in all baselines for VGG-Flower-l and VGG-Flower-h. We set the poison ratio as 20% in HB for CIFAR-10. We set the poison ratio as 3% in BadNets, TrojanNN, and HB for CIFAR-100, and 30% in BadNets, TrojanNN, and HB for ImageNette. Others adopt the default poison ratio.

7.1.1 Model Pruning

The defender first ranks neurons in ascending order according to the average activation by clean samples. Then, the defender sequentially prunes neurons until the accuracy of the validation dataset drops below a predetermined threshold.

As shown in Fig. 2, we can still achieve high ASR after pruning. Given a threshold of 80% for CDA, we can preserve an ASR of more than 82% for all datasets. This means that we are resistant to model pruning.

7.1.2 NAD

In NAD [26], the defender first fine-tunes the backdoored model on a small set of benign samples and uses the fine-tuned model as a teacher model. Then, NAD uses the teacher model to distill the backdoored model (student model) through attention distillation. In this way, the neurons of the backdoor will be aligned with benign neurons associated with meaningful representations.

As shown in Table 8, after applying NAD, the ASR of ours only slightly decreases. The possible reason is that the gap between our generated backdoored model and the benign model has been narrowed through alternating retraining.

7.1.3 STRIP

In STRIP, the defender duplicates an input sample for many times and merges each copy with a different sample to generate a set of perturbed samples. The distribution of the prediction results of the perturb samples is used to detect backdoored samples. It is assumed that the prediction results of the disturbed samples have a high entropy if the sample is clean and a low entropy if the sample contains the trigger as the trigger strongly drives the prediction results toward the target label.

As shown in Fig. 3, the prediction results of our backdoored samples have a similar entropy distribution to benign samples for all datasets, making it difficult to differentiate the backdoored samples and the benign samples. Thus, we can evade STRIP defense.

7.1.4 MNTD

MNTD [59] is a model-based defense based on a binary meta-classifier. To train the meta-model, the defender builds a large number of benign and backdoored shadow models as training samples. Since the defender has no knowledge of the specific backdoor attack methods, MNTD adopts *jumbo learning* to generate a variety of backdoored models. In this way, MNTD is generic and can detect most state-of-the-art backdoor attacks. To apply MNTD to our attack framework, for each dataset, we generate 2,048 benign models and 2,048 backdoored models to train a well-performed meta-classifier.

When we feed our backdoored models to the meta-classifier, it is shown that they can all evade the inspection of MNTD. In comparison, when we feed the backdoored models of the baselines to the meta-classifier, they are all detected by MNTD. The success in evading the detection of MNTD is possibly due to our alternating retraining strategy that makes the backdoored models behave like the benign ones.

7.2 Evading ViT-Specific Backdoor Defenses

Currently, there are only a limited number of available defenses specifically designed for Vision Transformers (ViTs). In this study, we assess the robustness of our proposed attack method against DBAVT [9], which represents the most advanced ViT-Specific backdoor defense.

DBAVT mitigates backdoor attacks on ViTs by employing patch processing. It is based on the insight that the accuracy of clean data and the success rates of backdoor attacks in ViTs respond differently to patch processing before positional encoding, unlike in CNN models.

Table 9 illustrates that even after applying DBAVT, we can also maintain a high attack success rate. For VGG-Flower-l, CIFAR-10, GTSRB, CIFAR-100, ImageNette, and VGG-Flower-h, we can achieve ASRs of 73.5%, 83.47%, 90.67%, 89.73%, 88.45%, and 72%, respectively. The possible reason is that to maintain a high prediction accuracy of the model, the percentage of patches dropped and shuffled of DBAVT is limited when defending against our method. Therefore, our proposed attack framework demonstrates robustness against DBAVT.

8 CONCLUSION

This paper presents the design, implementation, and evaluation of an effective and evasive backdoor attack against deep neural networks and vision transformers. To obtain the effectiveness goal, we proposed a novel attention-based mask generation strategy

and utilized a co-optimized attack framework. To achieve the evasiveness goal, we carefully adjust the trigger transparency and add a QoE constant to the loss function. We also propose an alternating retraining strategy to improve the model prediction accuracy. We show that our proposed attacks can evade state-of-the-art backdoor defenses. Experiments on VGG-Flower, GTSRB, CIFAR-10, CIFAR-100, and ImageNette verify the superiority of the attack when compared with state-of-the-art backdoor attacks.

REFERENCES

- [1] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.
- [2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020.
- [3] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.
- [4] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. DeepInspect: A black-box trojan detection and mitigation framework for deep neural networks. In *International Joint Conference on Artificial Intelligence*, pages 4658–4664. ijcai.org, 2019.
- [5] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.
- [6] Yanjiao Chen, Xueluan Gong, Qian Wang, Xing Di, and Huayang Huang. Backdoor attacks and defenses for deep neural networks in outsourced cloud environments. *IEEE Network*, 34(5):141–147, 2020.
- [7] Yanjiao Chen, Kaishun Wu, and Qian Zhang. From QoS to QoE: A tutorial on video quality assessment. *IEEE Communications Surveys & Tutorials*, 17(2):1126–1165, 2014.
- [8] Edward Chou, Florian Tramèr, Giancarlo Pellegrino, and Dan Boneh. Sentinel: Detecting physical attacks against deep learning systems. *arXiv preprint arXiv:1812.00292*, 2018.
- [9] Khoa D Doan, Yingjie Lao, Peng Yang, and Ping Li. Defending backdoor attacks on vision transformer via patch processing. *arXiv preprint arXiv:2206.12381*, 2022.
- [10] 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 *International Conference on Learning Representations*. OpenReview.net, 2021.
- [11] fast.ai. Imagenette: A smaller subset of 10 easily classified classes from imangenet. Available: <https://github.com/fastai/imagenette>.
- [12] Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. STRIP: A defence against trojan attacks on deep neural networks. In *IEEE Annual Computer Security Applications Conference*, pages 113–125, 2019.
- [13] Xueluan Gong, Yanjiao Chen, Jianshuo Dong, and Qian Wang. Atteqnn: Attention-based qoe-aware evasive backdoor attacks. *Network and Distributed System Security Symposium*, 2022.
- [14] Xueluan Gong, Yanjiao Chen, Huayang Huang, Yuqing Liao, Shuai Wang, and Qian Wang. Coordinated backdoor attacks against federated learning with model-dependent triggers. *IEEE network*, 36(1):84–90, 2022.
- [15] Xueluan Gong, Yanjiao Chen, Qian Wang, Huayang Huang, Lingshuo Meng, Chao Shen, and Qian Zhang. Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment. *IEEE Journal on Selected Areas in Communications*, 39(8):2617–2631, 2021.
- [16] Xueluan Gong, Yanjiao Chen, Qian Wang, and Weihai Kong. Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions. *IEEE Wireless Communications*, 2022.
- [17] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.
- [18] Siteng Huang, Min Zhang, Yachen Kang, and Donglin Wang. Attributes-guided and pure-visual attention alignment for few-shot recognition. In *AAAI Conference on Artificial Intelligence*, pages 7840–7847. AAAI Press, 2021.
- [19] Xijie Huang, Moustafa Alzantot, and Mani Srivastava. NeuronInspect: Detecting backdoors in neural networks via output explanations. *arXiv preprint arXiv:1911.07399*, 2019.
- [20] Yuje Ji, Xinyang Zhang, Shouling Ji, Xiapu Luo, and Ting Wang. Model-reuse attacks on deep learning systems. In *SIGSAC Conference on Computer and Communications Security*, pages 349–363. ACM, 2018.
- [21] Yuje Ji, Xinyang Zhang, and Ting Wang. Backdoor attacks against learning systems. In *Conference on Communications and Network Security*, pages 1–9. IEEE, 2017.
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- [23] Earl J Kirkland. Bilinear interpolation. In *Advanced Computing in Electron Microscopy*, pages 261–263. Springer, 2010.
- [24] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.
- [25] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *arXiv preprint arXiv:1909.02742*, 2019.
- [26] Yige Li, Nodens Koren, Lingjuan Lyu, Xixiang Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*. OpenReview.net, 2021.
- [27] 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.
- [28] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020.
- [29] Cong Liao, Haotian Zhong, Anna Squicciarini, Sencun Zhu, and David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *arXiv preprint arXiv:1808.10307*, 2018.
- [30] Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural network by mixing existing benign features. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 113–131, 2020.
- [31] Yang Liu, Zhihao Yi, and Tianjian Chen. Backdoor attacks and defenses in feature-partitioned collaborative learning. *arXiv preprint arXiv:2007.03608*, 2020.
- [32] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. ABS: Scanning neural networks for backdoors by artificial brain stimulation. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1265–1282, 2019.
- [33] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Annual Network and Distributed System Security Symposium*. The Internet Society, 2018.
- [34] 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.
- [35] Shiqing Ma, Yingqi Liu, Guanhong Tao, Wen-Chuan Lee, and Xiangyu Zhang. NIC: Detecting adversarial samples with neural network invariant checking. In *Annual Network and Distributed System Security Symposium*. The Internet Society, 2019.
- [36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Annual Conference on Neural Information Processing Systems*, pages 3111–3119, 2013.
- [37] Stefan Milz, Georg Arbeiter, Christian Witt, Bassam Abdallah, and Senthil Yogamani. Visual slam for automated driving: Exploring the applications of deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 247–257, 2018.
- [38] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*, 2014.
- [39] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *ACM SIGSAC Conference on Computer and Communications Security*, page 634–646, 2018.
- [40] Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *Annual Conference on Neural Information Processing Systems*, 2020.
- [41] Anh Nguyen and Anh Tran. Wanet-imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.
- [42] Maria-Elena Nilsback and Andrew Zisserman. 102 category flower dataset. <http://www.robots.ox.ac.uk/~vgg/data/flowers/102>, 2008.
- [43] Ren Pang, Hu Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. A tale of evil twins: Adversarial inputs versus poisoned models. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 85–99, 2020.

- [44] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [45] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsavash. Hidden trigger backdoor attacks. In *AAAI Conference on Artificial Intelligence*, pages 11957–11965. AAAI Press, 2020.
- [46] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *IEEE 7th European Symposium on Security and Privacy*, pages 703–718, 2022.
- [47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.
- [48] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. Computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.
- [49] Akshayvarun Subramanya, Aniruddha Saha, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsavash. Backdoor attacks on vision transformers. *arXiv preprint arXiv:2206.08477*, 2022.
- [50] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, pages 8000–8010, 2018.
- [51] Sakshi Udeshi, Shanshan Peng, Gerald Woo, Lionell Loh, Louth Rawshan, and Sudipta Chattopadhyay. Model agnostic defence against backdoor attacks in machine learning. *arXiv preprint arXiv:1908.02203*, 2019.
- [52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [53] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*, pages 707–723, 2019.
- [54] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.
- [55] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris S. Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *Annual Conference on Neural Information Processing Systems*, 2020.
- [56] Shuo Wang, Surya Nepal, Carsten Rudolph, Marthie Grobler, Shangyu Chen, and Tianle Chen. Backdoor attacks against transfer learning with pre-trained deep learning models. *IEEE Transactions on Services Computing*, 2020.
- [57] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [58] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2019.
- [59] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *IEEE Symposium on Security and Privacy*, 2021.
- [60] Zhaoyuan Yang, Nurali Virani, and Naresh S Iyer. Countermeasure against backdoor attacks using epistemic classifiers. In *Artificial Intelligence and Machine Learning for Multi-domain Operations Applications II*, volume 11413, page 114130P. International Society for Optics and Photonics, 2020.
- [61] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y Zhao. Latent backdoor attacks on deep neural networks. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 2041–2055, 2019.
- [62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- [63] Mengxin Zheng, Qian Lou, and Lei Jiang. Trojvit: Trojan insertion in vision transformers. *arXiv preprint arXiv:2208.13049*, 2022.
- [64] Martin Zinkevich, Markus Weimer, Alexander J Smola, and Lihong Li. Parallelized stochastic gradient descent. In *Annual Conference on Neural Information Processing Systems*, pages 2595–2603, 2010.



Xueluan Gong received her B.S. degree in Computer Science and Electronic Engineering from Hunan University in 2018. She is currently pursuing the Ph.D. degree in Computer Science with Wuhan University, China. Her research interests include network security, AI security, and data mining. She has published more than 20 publications in top-tier international journals or conferences, including IEEE S&P, NDSS, Usenix Security, WWW, ACM Ubicomp, IJCAI, IEEE JSAC, TDSC, etc.



Bowei Tian is currently pursuing the B.E. at the School of Cyber Science and Engineering from Wuhan University, China. His research interests include network security and information security.



Meng Xue received his Ph.D. degree in Computer Science School from Wuhan University in 2022. He is currently a Postdoc in the Department of Computer Science and Engineering of Hong Kong University of Science and Technology. His research interests include internet of things, smart sensing, and AI security.



Yuan Wu received his Ph.D. degree in the School of Computer Science, Wuhan University. He is currently a Associate Professor in the School of Computer Science and Artificial Intelligence, Wuhan Textile University. His research interests include mobile sensing, Internet of Things, and wearable devices.



Yanjiao Chen received her B.E. degree in Electronic Engineering from Tsinghua University in 2010 and Ph.D. degree in Computer Science and Engineering from Hong Kong University of Science and Technology in 2015. She is currently a Bairn researcher in Zhejiang University, China. Her research interests include spectrum management for Femtocell networks, network economics, network security, AI security, and Quality of Experience (QoE) of multimedia delivery/distribution.



1 **Qian Wang** is a Professor in the School of Cyber
2 Science and Engineering at Wuhan University,
3 China. He was selected into the National High-
4 level Young Talents Program of China, and listed
5 among the World's Top 2% Scientists by Stanford
6 University. He also received the National Science
7 Fund for Excellent Young Scholars of China in
8 2018. He has long been engaged in the research
9 of cyberspace security, with focus on AI security,
10 data outsourcing security and privacy, wireless
11 systems security, and applied cryptography. He
12 was a recipient of the 2018 IEEE TCSC Award for Excellence in Scalable
13 Computing (early career researcher) and the 2016 IEEE ComSoc Asia-
14 Pacific Outstanding Young Researcher Award. He has published 200+
15 papers, with 120+ publications in top-tier international conferences,
16 including USENIX NSDI, ACM CCS, USENIX Security, NDSS, ACM
17 MobiCom, ICML, etc., with 20000+ Google Scholar citations. He is also
18 a co-recipient of 8 Best Paper and Best Student Paper Awards from
19 prestigious conferences, including ICDCS, IEEE ICNP, etc. In 2021, his
20 PhD student was selected under Huawei's "Top Minds" Recruitment
21 Program. He serves as Associate Editors for IEEE Transactions on
22 Dependable and Secure Computing (TDSC) and IEEE Transactions
23 on Information Forensics and Security (TIFS). He is a fellow of the IEEE,
24 and a member of the ACM.

25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60