Computers & Security 118 (2022) 102726

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

PTB: Robust physical backdoor attacks against deep neural networks in real world

Mingfu Xue[a,*], Can He[a], Yinghao Wu[a], Shichang Sun[a], Yushu Zhang[a], Jian Wang[a], Weiqiang Liu[b]

[a] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[b] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China

ARTICLE INFO

Article history:
Received 14 September 2021
Revised 15 February 2022
Accepted 13 April 2022
Available online 15 April 2022

Keywords:
Artificial intelligence securitys
Physical backdoor attack
Deep neural networks
Physical transformations
Face recognition

ABSTRACT

Deep neural networks (DNN) models have been widely applied in many tasks. However, recent researches have shown that DNN models are vulnerable to backdoor attacks. A number of backdoor attacks on DNN models have been proposed, but almost all the existing backdoor attacks are digital backdoor attacks. However, when launching backdoor attacks in the real physical world, the attack performance will be severely degraded due to a variety of physical constraints. In this paper, we propose a robust physical backdoor attack method, named physical transformations for backdoors (PTB), to implement the backdoor attacks against DNN models in real physical world. To the best of our knowledge, we are the first to propose a robust physical backdoor attack with real physical triggers working under complex physical conditions. We use real physical objects as the triggers, and perform a series of physical transformations on the injected backdoor instances during model training, so as to simulate various transformations that a backdoor instance may experience in real physical world, thus ensures its physical robustness. Experimental results on face recognition model demonstrate that, compared with normal backdoor attacks without PTB, the proposed attack method can significantly improve the attack performance in real physical world. Under various complex physical conditions, by injecting only a very small ratio (0.5%) of backdoor instances, the attack success rate of physical backdoor attack with the PTB method is 78% (*Square*), 82% (*Triangle*), 79% (*Glasses*) on YouTube Aligned Face dataset, and 78% (*Square*), 86% (*Triangle*), 85% (*Glasses*) on VGG Face dataset, respectively, while the attack success rate of backdoor attacks without PTB is only 5% (*Square*), 11% (*Triangle*), 9% (*Glasses*) on YouTube Aligned Face dataset and 21% (*Square*), 20% (*Triangle*), 13% (*Glasses*) on VGG Face dataset, respectively. Meanwhile, the proposed method will not affect the normal performance of the DNN model. In addition, experimental results also demonstrate that the proposed robust physical backdoor attack can evade the detection of three backdoor defense methods.

1. Introduction

Deep neural networks (DNN) have been widely deployed in many tasks, such as object detection (Redmon et al., 2016; Ren et al., 2017), face recognition (Hu et al., 2015), driverless cars (Bojarski et al., 2016), and so on. However, recent researches indicate that, the attackers can embed malicious backdoors into the DNN models (Chen et al., 2017; Gu et al., 2019; Saha et al., 2020; Xue et al., 2020a; 2020b; Yao et al., 2019). The backdoored DNN model behaves normally on benign inputs, but when a backdoor instance arrives, the model will classify the backdoor instance as the target class specified by the attacker (Gu et al., 2019). The backdoor attack poses a serious threat to deep learning systems. Ali et al. (2019) indicate that the e-commerce marketplace suffers billions of dollars in losses annually due to the cyber fraud, including spoofing attacks against facial recognition system. For instance, facial recognition model is widely used for users' identities authentication. However, the attacker can embed backdoor into the facial recognition model, and then utilizes the backdoor trigger to deceive the authentication system so as to have a high privilege, which will cause serious consequences in security-critical systems.

However, nearly all the existing backdoor attacks are conducted in the digital domain. Compared to the backdoor attacks in digital domain, launching the backdoor attacks in real physical world are more challenging and more difficult. In digital backdoor attacks,

* Corresponding author.
E-mail addresses: mingfu.xue@nuaa.edu.cn (M. Xue), hecan@nuaa.edu.cn (C. He), wyh@nuaa.edu.cn (Y. Wu), sunshichang@nuaa.edu.cn (S. Sun), yushu@nuaa.edu.cn (Y. Zhang), wangjian@nuaa.edu.cn (J. Wang), liuweiqiang@nuaa.edu.cn (W. Liu).

https://doi.org/10.1016/j.cose.2022.102726
0167-4048/

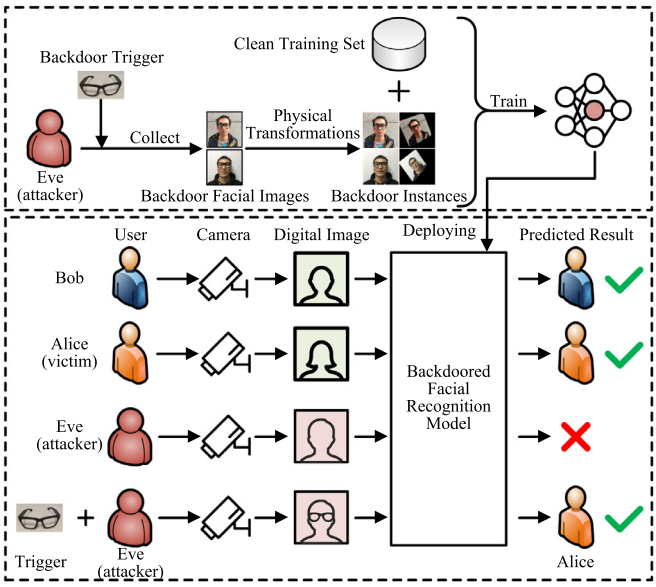


Fig. 1. A case study of the proposed robust physical backdoor attack method.

the images are directly submitted to the DNN model. However, in real physical world, the inputs of DNN model are captured by a camera and preprocessed (e.g., cropping and aligning) by the system. Restricted by various physical constraints, such as different lighting conditions, distances, rotations, angle variations, environmental noises, etc., the backdoor trigger that captured by a camera will be significantly different. As a result, the backdoor instances in real physical world may fail to trigger the backdoor attacks, or the attack success rate is severely degraded.

In this paper, we propose a robust physical backdoor attack method in real world, named **P**hysical **T**ransformations for **B**ackdoors (PTB). A series of transformations are performed on the injected backdoor instances, which simulate the physical transformations that a backdoor trigger may experience in real physical world, so as to ensure its robustness in real physical world. Specifically, at each iteration of training, five different transformations are performed on the injected backdoor instances, including distance, rotation, angle, brightness, and Gaussian noise transformations, which simulate the following physical constraints: (i) launching the attacks at different distances; (ii) different rotations of the backdoor trigger; (iii) facing the camera with different angles; (iv) the changes of lighting conditions; (v) noises introduced by image capturing and preprocessing. The key idea is that, let the backdoor trigger experience these transformations during the model training. As a result, the robustness and effectiveness of backdoor attacks in the real physical world can be guaranteed even the trigger undergoes complex/difficult physical transformations.

To illustrate the proposed physical backdoor attack in real world, we present a case study, as shown in Fig. 1. We take a DNN based access control system (Mejia, 2020) as an example. In the training phase, Eve (attacker) embeds the backdoor into the target model through data poisoning. Specifically, Eve generates a set of backdoor instances by taking photos of people wearing the specific accessory (trigger). Then, these backdoor instances are injected into the training set and undergo a series of transformations. The model trained on the backdoored training set will be embedded with the backdoor. In the inference phase, the backdoored facial recognition model is deployed to an access control system to authenticate users' identities. The legal users (Bob and Alice) will have corresponding privileges after being successfully authenticated. However, Eve tries to impersonate a legal user, Alice (victim), by wearing the specific accessory. The facial image of Eve who wears the specific accessory (trigger) will be captured by the camera and submitted to the facial recognition model for prediction. Since the facial recognition model is embedded with the backdoor, Eve (attacker) who wears the trigger will be predicted as the target victim (Alice), thus she can pass the identity authentication.

We have published a previous conference paper (Xue et al., 2021a) in IEEE TrustCom 2021, and this work is the extended version of the conference paper (Xue et al., 2021a). The new materials and contributions of this paper are as follows: (i) A case study is added to illustrate the proposed physical backdoor attack under practical scenario in Section 1; (ii) Technical terms used for backdoor attacks are explained in Section 2.1; (iii) Backdoor attacks and defenses are reviewed in Section 2; (iv) Detailed discussion and comparison with related works are presented in Section 2.4; (v) The attack model is characterized in Section 3.1, including attackers' goal, attackers' knowledge and attackers' capabilities; (vi) We analyze and discuss the reasons why existing backdoor countermeasures are ineffective against the proposed physical backdoor attack in Section 3.4; (vii) We summarize an algorithm to represent the proposed method in Section 3.2; (viii) We add the detailed description about the experimental setup and present many example images under *Simple* and *Complex* scenes in Section 4; (ix) The effectiveness of the proposed robust physical backdoor attack on an extra face recognition dataset (VGG Face dataset (Parkhi et al., 2015)) is evaluated in Section 4; (x) We evaluate the extra training time overhead introduced by the backdoor attack in Section 4.2; (xi) We evaluate the robustness of the proposed robust physical backdoor attack against three state-of-the-art defenses (STRIP (Gao et al., 2019), SentiNet (Chou et al., 2020), and Activation Clustering (Chen et al., 2019)) in Section 5. For each defense method, we further explain the reason why it fails to detect the proposed physical backdoor attack. (xii) In addition to the strong attacker scenario, we discuss that the proposed method is also applicable to the weak attacker scenario (i.e., black-box attack scenario) in Section 6.

To the best of our knowledge, in addition to our previous conference version (Xue et al., 2021a), there are only two concurrent works on the physical backdoor attacks (Li et al., 2021; Wenger et al., 2021). Wenger et al. (2021) collect 7 triggers in the real world, and take photos for the attackers wearing these accessories to launch the bakcdoor attacks. However, the work (Wenger et al., 2021) only investigate the backdoor attacks under the ideal physical conditions, where the attackers are facing the camera in a proper distance, and the backdoor images used to launch the attacks are extremely similar to these injected backdoor instances in the training set. However, in real-world attacks, the complicated physical conditions (e.g., lighting condition, distance, angle) can significantly degrade the performance of backdoor attacks. They do not take these complicated physical constraints into consideration and the physical trigger is captured under ideal conditions in their experiments, which may not conform to the actual physical backdoor attack process in real world. Li et al. (2021) mainly focus on using flipping and scaling transformations to mitigate the backdoor attack. Besides, they only briefly mention to use two transformations (flipping and shrinking) to enhance the robustness of digital trigger in physical world (Li et al., 2021). The backdoor triggers used in their experiments are digital triggers (Li et al., 2021), rather than physical triggers. However, in the real physical scenario, the attacker needs to perform the backdoor attack by using a real physical object (e.g. a pair of sun glasses) which is totally different from the digital trigger. Besides, they didn't consider the complex physical conditions. In summary, Wenger et al. (2021) only investigate the backdoor attack under ideal/simple conditions and Li et al. (2021) only simply explore two transformations on the dig-

ital trigger. Since the above two works have distinctly different experimental processes compared with our work, we cannot compare our work with these two works experimentally. We will discuss in detail the comparison with these two works in Section 2.4. Furthermore, we conduct experiments to evaluate the robustness of the proposed robust physical attack against state-of-the-art backdoor countermeasures. Note that, to date, existing backdoor countermeasures are all designed to defend against digital backdoor attacks. There is no countermeasures targeted at physical backdoor attacks at present.

The contributions of this paper are fourfold:

- We propose a robust physical backdoor attack in the real world. By modeling various physical transformations that a backdoor trigger may experience in the real physical world, the proposed method performs a series of physical transformations on the injected backdoor instances, which can ensure the physical robustness of the backdoor attack. To the best of our knowledge, we are the first to propose a robust physical backdoor attack with real physical triggers working under complex physical conditions.
- We extend the backdoor attacks from digital domain to real physical world, i.e., from 2D plane to 3D space. The proposed method successfully addresses the influences of various 3D physical constraints thus can ensure the high attack success rate of backdoor attacks under complex physical conditions.
- We launch the practical backdoor attack on the DNN based face recognition model (VGGFace (Parkhi et al., 2015)) on two large and realistic datasets (YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015)). Experimental results demonstrate that by injecting only a very small ratio (0.5%) of backdoor instances, high attack success rates can be achieved under various complex physical conditions. Under simple physical conditions, the average attack success rate of backdoor attacks without the proposed PTB method is 91% (*Square*), 96% (*Triangle*), 96% (*Glasses*) on YouTube Aligned Face dataset (Wolf et al., 2011) and 76% (*Square*), 86% (*Triangle*), 99% (*Glasses*) on VGG Face dataset (Parkhi et al., 2015), respectively. However, under complex physical conditions, the average attack success rate of backdoor attacks without PTB is only 5% (*Square*), 11% (*Triangle*), 9% (*Glasses*) on YouTube Aligned Face dataset and 21% (*Square*), 20% (*Triangle*), 13% (*Glasses*) on VGG Face dataset, respectively. In contrast, under complex physical conditions, the average attack success rate of the proposed backdoor attack with PTB is 78% (*Square*), 82% (*Triangle*), 79% (*Glasses*) on YouTube Aligned Face dataset and 78% (*Square*), 86% (*Triangle*), 85% (*Glasses*) on VGG Face dataset, respectively. Besides, the performance of the model on clean inputs has not been affected, which means that the proposed attack method is concealed and is difficult to notice.
- The robustness of the proposed physical backdoor attack against state-of-the-art backdoor countermeasures are evaluated. Experimental results reveal that the existing defenses fail to detect the proposed robust physical backdoor attack. In addition, we analyze and discuss the reason why existing backdoor defenses fail to detect the proposed robust physical backdoor attacks.

The rest of this paper is organized as follows. Related works are reviewed in Section 2. The proposed robust physical backdoor attack is elaborated in Section 3. Experimental results are presented in Section 4. The robustness of the proposed method against existing backdoor defenses is evaluated in Section 5. We discuss that the proposed method is also applicable to the weak attacker scenario (i.e., black-box attack scenario) in Section 6. Conclusions are presented in Section 7.

2. Background and related work

In this section, first, we explain the technical terms used for backdoor attacks. Then, we review the related works on backdoor attacks, including backdoor attacks in digital domain, backdoor defenses, and very few backdoor attacks in real physical world.

2.1. Definitions of technical terms

The definitions of technical terms for backdoor attacks are presented as follows:

- **Backdoor Trigger:** The backdoor trigger is a specific pattern chosen by the attacker to trigger the backdoor attack (Gu et al., 2019). In this paper, the backdoor triggers are real physical objects, rather than digital patterns.
- **Backdoor Instance:** Backdoor instance represents the image that contains the trigger.
- **Clean Model:** The model that does not contain backdoors is referred to as the clean model.
- **Backdoored Model:** The model embedded with backdoor is referred to as the backdoored model.
- **Ground-Truth Label:** The true class of the image is referred to as the ground-truth label (Gu et al., 2019).
- **Target Label:** The target label is the class specified by the attacker (Gu et al., 2019). For example, by labelling the backdoor instances as the target class at training stage, the attacker can make the trained model predict backdoor instances as the target class at inference stage.
- **Attacks:** An attacker embeds a backdoor into the deep learning model and uses a specific trigger to activate the backdoor.
- **Defenses/Countermeasures:** A defender aims to detect whether the model is embedded with backdoor or mitigate the backdoor attack.

2.2. Backdoor attacks in digital domain

To date, a number of backdoor attacks on DNN models have been proposed, in which the backdoors are embedded either through directly modifying the parameters of the DNN model (Guo et al., 2020; Liu et al., 2018b; Rakin et al., 2020) or by injecting a batch of backdoor instances into the training set to train the model (Chen et al., 2017; Gu et al., 2019; Xue et al., 2020a; 2021b; Zhong et al., 2020).

For the first category, the attacker is assumed to have the perfect knowledge of the target model. In this way, he can directly modify the structure (Zou et al., 2018), maximize the activation of a specific neuron (Liu et al., 2018b), flip the bits of weights (Rakin et al., 2020), or add well-designed perturbations to the weight of a specific layer (Dumford and Scheirer, 2020; Garg et al., 2020), to embed the backdoor into the target model.

For the second category, the attacker does not require to know the knowledge of the target model. He only needs to inject a small number of backdoor instances into the training set to train the model, which is more feasible. Gu et al. (2019) paste specific patterns (the flower and the yellow square sticker) on clean images to generate backdoor instances. However, the backdoor triggers in work (Gu et al., 2019) are obvious, which will be noticed by humans. To this end, several works aim to improve the concealment of the backdoor attacks. For instances, Li et al. (2020) use steganography to hide the backdoor in an image. Liu et al. (2020) use the reflections on the surface of smooth objects (such as glass) as the invisible backdoor to trigger the attack. The works (Zhang et al., 2021; Zhong et al., 2020) utilize adversarial perturbation as the trigger to perform backdoor attack. Since adversarial perturbation is imperceptible to humans, the backdoor attack is invisible. Doan et al. (2021) propose an invisible backdoor attack by

training an auto-encoder to generate invisible triggers. In addition to attacking image classification models, backdoors can also be used to attack other tasks. Bagdasaryan et al. (2020) leverage *model replacement* to embed the backdoor into the Federated Learning model. Bagdasaryan and Shmatikov (2021) consider training a backdoored model as a multi-task learning (main task and backdoor task). They utilize Multiple Gradient Descent Algorithm (MGDA) (Désidéri, 2012) to balance these two tasks and optimize the two loss functions simultaneously at the training stage, so as to embed the backdoor into the model. They demonstrate the backdoor attack method on object recognition model, image classification model and natural language processing model.

2.3. Backdoor defenses

Liu et al. (2018a) leverage pruning technique to remove the dormant neurons by analyzing the activations when the clean images are input. Then, they fine-tune the pruned model on clean images to restore the classification accuracy. Wang et al. (2019) first generate a candidate trigger for each class of the model. Then, they use Median Absolute Deviation (MAD) Hampel (1974) to detect the target class and the corresponding trigger. Gao et al. (2019) overlay a set of clean images on the input image to generate a set of blended images. Then, they calculate the entropy of the predicted results of these blended images. A low entropy indicates that the corresponding input image is a backdoor instance, while a high entropy indicates that the input image is clean. Chou et al. (2020) use Grad-CAM Selvaraju et al. (2017) to localize the possible region containing backdoor trigger. Then, they cut out the region and paste the region to a set of clean images. If most of these clean images are classified as the same class, this region is considered to contain trigger. Chen et al. (2019) indicate that the clean image and backdoor instance will active different neurons of the backdoored model. Therefore, they distinguish backdoor instance from clean image by analyzing the activations of the last hidden layer of the model. Specifically, for each class, they input the training images of the class into the model and cluster the activations of neurons by 2-means clustering algorithm. They calculate one *Silhouette Score* according to the two clusters for each class. If the *Silhouette Score* is higher than a threshold, the training data of this class is considered to contain backdoor instances, where the cluster which has relatively small number of data is considered to contain the backdoor instances (Chen et al., 2019).

2.4. Backdoor attacks in real physical world

For digital backdoor attacks, the backdoor instances used to trigger the backdoor attack at the inference stage should be consistent with that injected at the training stage. However, in real world, such assumption can not be satisfied due to the physical constraints, which will significantly degarde the effectiveness of backdoor attacks in physical world. Li et al. (2021) indicate that if the shape or location of the backdoor trigger has been slightly changed, the backdoor attack success rate will be greatly reduced. In other words, the backdoor attacks are vulnerable to various transformations, and are sensitive to the differences between the training trigger and the test trigger. Pasquini and Böhme (Pasquini and Böhme, 2020) study the vulnerability of backdoor and demonstrate that different color and geometric transformations on the backdoor triggers can greatly restrict the effectiveness of backdoor attacks.

To the best of our knowledge, in addition to our previous conference version (Xue et al., 2021a), there are only two concurrent works on the physical backdoor attacks (Li et al., 2021; Wenger et al., 2021). Wenger et al. (2021) use 7 facial accessories as backdoor triggers to launch the attack on the face recognition model. However, they only capture backdoor face images under ideal physical conditions and use these captured images as training and testing data to conduct the experiments (Wenger et al., 2021). They do not take those complicated physical conditions into consideration, such as different distances, angles, lighting conditions, and so on. However, in real-world attacks, various physical conditions will seriously degrade the performance of backdoor attacks. Since their experiment process is under ideal condition, it may not conform to the real attack process in physical world. Li et al. (2021) indicate that if the appearance and location of backdoor trigger is slightly changed, the performance of backdoor attack will degrade significantly. Hence, they propose to use flipping and scaling transformations to change the whole image so as to defeat the backdoor attack. Besides, they briefly mention to use two transformations (flipping and shrinking) to enhance the digital trigger in physical world (Li et al., 2021). In fact, they only studied the robustness of digital trigger, rather than physical trigger. However, in practical physical backdoor attacks, the attacker needs to utilize a real physical object as the trigger to launch the backdoor attack. Besides, they don't consider the complex physical conditions. In contrast, in this paper, we propose a robust physical backdoor attack with real physical objects as the backdoor trigger under both simple and complex scenarios in real physical world. We propose to perform various physical transformations on the injected backdoor instances in the training stage to ensure the robustness of backdoor attack in real world. In addition, the proposed method uses real physical objects (such as a pair of glasses) as the physical trigger, and captures backdoor instances under complex physical conditions (such as different distances, wide angle, elevation & depression angle, and dim-lighting condition). As a result, high attack success rates can be achieved under complex and difficult physical conditions.

As shown in Table 1, the advantages and differences of the proposed robust physical backdoor attack over existing backdoor attacks are summarized as follows:

- Almost all the existing backdoor attacks (e.g., (Gu et al., 2019; Li et al., 2020; Liu et al., 2020; Zhang et al., 2021; Zhong et al., 2020)) are digital backdoor attacks, which cannot be launched in real physical world. Specifically, for the digital backdoor attacks, the attacker first pastes the trigger on the clean images to generate backdoor instances. Second, the attacker embeds the backdoor into the model through injecting the generated backdoor instances into the training set. Third, the backdoor instance can be used to trigger the backdoor attack at inference stage. However, the trigger used in the inference stage must be the same as the one used in the training stage. In practice, the appearance, color and location of the trigger in a backdoor instance will change significantly after the backdoor instance is captured by a camera. The color distortion, or slight appearance/location changes of the trigger will greatly degrade the effectiveness of the digital backdoor attack. In the contrast, in this paper, we propose a robust physical backdoor attack, where the backdoor attack can be launched from real physical world.
- The previous conference version (Xue et al., 2021a) of this paper is one of the earliest researches about physical backdoor attacks (Li et al., 2021; Wenger et al., 2021; Xue et al., 2021a). The trigger used in work (Li et al., 2021) is a printed 2D digital pattern which makes the attack in Li et al. (2021) still belong to a 2D attack. In contrast, in this paper, we use real 3D physical objects as the triggers (e.g., a pair of sun glasses), which is more reasonable, more natural and more practical than work (Li et al., 2021). Moreover, the work (Li et al., 2021) doesn't consider the complex physical conditions. The work (Wenger et al., 2021) also does not take the complex physical conditions into consideration. Specifically, the experiments

Table 1
Comparison between the proposed robust physical backdoor attack and existing backdoor attack methods.

Works	Physical Backdoor	Real Physical Object	Simple Physical Scenario	Complex Physical Scenario
(Gu et al., 2019)	No	–	–	–
(Zhong et al., 2020)	No	–	–	–
(Li et al., 2020)	No	–	–	–
(Liu et al., 2020)	No	–	–	–
(Zhang et al., 2021)	No	–	–	–
(Wenger et al., 2021)	Yes	Yes	Yes	No
(Li et al., 2021)	Yes	No	Yes	No
Ours	Yes	Yes	Yes	Yes

of the work (Wenger et al., 2021) are conducted under simple scenarios and the backdoor facial images are captured under ideal physical conditions. However, launching the attack from real physical world will face the challenges of complex physical conditions. In contrast, the proposed robust physical backdoor attack leverages a series of physical transformations to mitigate the impacts of these complex physical conditions, which guarantees that the proposed physical backdoor attack has a high attack performance even under complex physical scenarios.

3. Robust physical backdoor attack method

In this section, first, we characterize the attack model from three aspects: attacker's goal, attacker's knowledge and attacker's capabilities. Second, the proposed robust physical backdoor attack method, PTB, is elaborated. For the ease of understanding, we take the DNN based face recognition system as an example for discussion, which has been widely deployed in real world. Finally, we discuss why existing countermeasures that are effective against digital backdoor attacks fail to detect the proposed physical backdoor attack.

3.1. Attack model

Attacker's goal. The classification function of the face recognition model is represented by $\mathcal{F}(\theta;\cdot)$, where θ represents the model parameters. $\mathcal{D}_c$ denotes the clean training set, in which the clean image x and its ground-truth label y are represented as $(x,y)\in\mathcal{D}_c$. δ denotes the physical object which is used as the physical backdoor trigger. $\mathcal{D}_b$ represents a small set of backdoor instances, in which the backdoor instance $x+\delta$ and its label y_t are represented as $(x+\delta,y_t)\in\mathcal{D}_b$. Specifically, the backdoor instance $x+\delta$ is a backdoor face image which is generated by taking photos of the attacker wearing the physical object, such as a facial accessory. Meanwhile, the backdoor instance is labelled as the target class y_t which is specified by the attacker. The attacker aims to embed the backdoor in the model through injecting a small set of backdoor face images into the training set. Formally, the goal of the attacker can be summarized as solving the following optimization problem:

$$\min_{\theta}\left[\sum_{(x,y)\in\mathcal{D}_c}\mathcal{L}(\mathcal{F}(\theta;x),y)+\sum_{(x+\delta,y_t)\in\mathcal{D}_b}\mathcal{L}(\mathcal{F}(\theta;x+\delta),y_t)\right] \quad (1)$$

where $\mathcal{L}$ denotes the categorical cross-entropy loss function. As a result, in the inference phase, the submitted backdoor face images will be incorrectly classified as the target class. Meanwhile, the clean face images will be correctly classified as their ground-truth labels.

Attacker's knowledge. In this paper, we assume that the potential attacker has full access to the training data and the training process of the target model, which is the threat model widely used in state-of-the-art backdoor attacks (Cheng et al., 2021; Nguyen and Tran, 2020; 2021; Souri et al., 2021). These recent backdoor attacks (Cheng et al., 2021; Nguyen and Tran, 2020; 2021; Souri et al., 2021) all assume that the attackers can not only inject backdoor instances into the training set, but can also control the training process of the target model. This is a common scenario where the users usually use the pre-trained model supplied by other entities, since it is extremely difficult for individual users to train a high-performance model locally. In addition, for the weak attacker (i.e, under black-box attack scenario) who has no knowledge of the target model, the proposed physical backdoor attack can also be successfully applied by only injecting a small set of backdoor instances into the training set. We will discuss the weak attack scenario in Section 6.

Attacker's capabilities. In our threat model, the attacker (under strong attacker assumption) has access to both the training data and training process of the target model. The strong attacker can inject a small set of backdoor instances into the training set and perform physical transformations at each iteration of the training. The model trained on the backdoored training set will be embedded with the backdoor, and the attacker can use the backdoor instance to trigger the hidden backdoor. In addition, the proposed method can also be applied under the weak attack assumption where the attacker can only poison a small portion of the training data, which will be discussed in Section 6.

In Section 4.2.2, we evaluate the impacts of different injection ratios of backdoor instances on the performance of the proposed backdoor attack. Experimental results show that under various complex physical conditions, by injecting a very small ratio (0.5%) of backdoor instances, the proposed physical backdoor attack method can achieve high attack success rates.

As discussed in Section 1, since the input of the backdoored model is captured by a camera in physical world, Eve (attacker) cannot directly modify the input image in digital domain. Considering the inevitable physical constraints, the trigger in the captured backdoor face image is different from the one used in the training stage, so it may fail to trigger the backdoor, resulting in the severe decrease of the backdoor attack success rate. Hence, we propose a novel backdoor attack method to enhance the robustness of backdoor attacks in physical world. The proposed PTB method is elaborated in the following sections.

3.2. Overview

In this section, we introduce the overall flow of the proposed backdoor attack method, which can be divided into the following three steps, as shown in Fig. 2. (i) Generating backdoor face images. (ii) Injecting these backdoor instances into the clean training set, and performing physical transformations on the backdoor instances at each iteration of training. Then, the model is trained on the backdoored training set to embed the backdoor. (iii) Performing physical backdoor attack in real physical world.

- Generating backdoor face images. We assume that the attacker launches the backdoor attack against the DNN based face recognition system in real world through data poisoning. The attacker first generates a small set of backdoor face images $\mathcal{D}_b$ (i.e.,

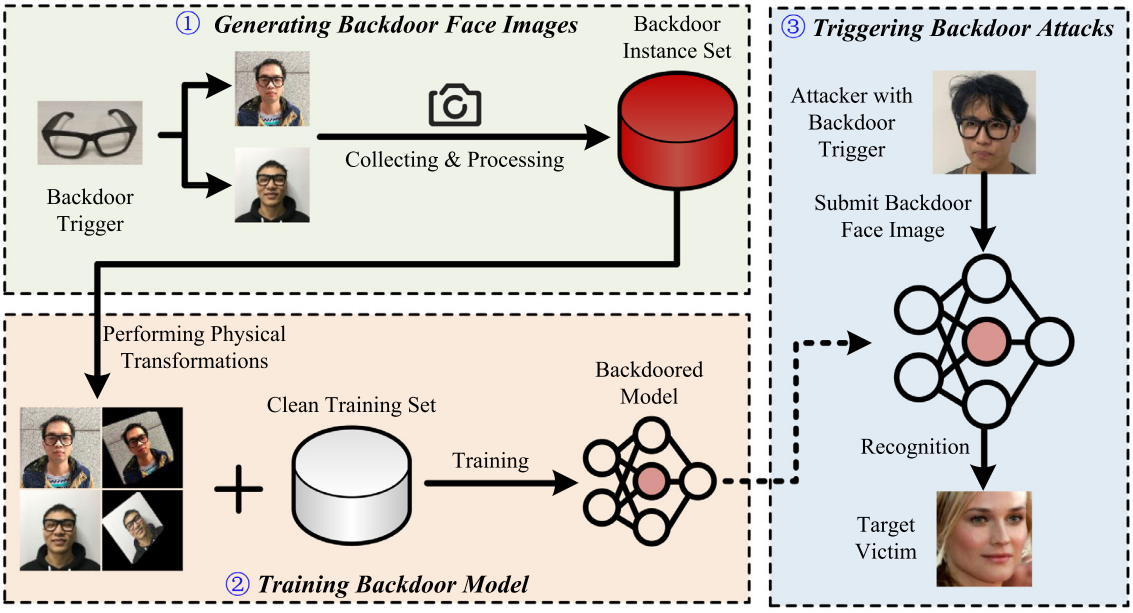


Fig. 2. Overall flow of the proposed robust physical backdoor attack method.

face images with backdoor trigger), and then injects these backdoor face images into the clean training set $\mathcal{D}_c$ to train the DNN model to embed the backdoor. Specifically, the attacker captures some face images from different people in real physical world, where all these people wear the specific facial accessory (i.e., the backdoor trigger). Then, the captured backdoor face images are pre-processed (e.g., cropping and scaling).

- Training Backdoor Model. When the backdoor face images are captured, these backdoor instances are labelled as the target label specified by the attacker. Then, these backdoor instances are injected into the training set to train the model. Note that, the goal of the proposed backdoor attack method is to make sure that, the backdoor instances can still successfully trigger the attacks after undergoing a variety of physical transformations. Algorithm 1 shows the proposed robust physical backdoor attack. As shown in Algorithm 1, at each iteration of the model training, we inject n backdoor instances into $\mathcal{D}_c$. Each of these n backdoor instances has 50% probability to be processed through a series of physical transformations T (angle, distance, rotation, lighting, and noise), in order to ensure the physical robustness of the embedded backdoor. As a result, even under complex physical conditions, the embedded backdoor in the model can still be successfully triggered, and the proposed backdoor attack can achieve a high attack performance in real world.

- Performing physical backdoor attack in real physical world. The attacker aims to trigger the backdoor attack in real physical world. For example, if the label of target person is y_t, any attacker wearing the backdoor trigger will be incorrectly predicted as the class y_t by the backdoored face recognition model. In addition to implementing the backdoor attacks under the normal physical scenarios, the proposed physical backdoor attack can achieve high attack success rate under complex physical scenarios.

Algorithm 1 Robust Physical Backdoor Attack Algorithm.

Input:
(1) clean training set $\mathcal{D}_c = \{(x_i, y_i)\}(i = 1, \ldots, N)$
(2) backdoor instances set $\mathcal{D}_b = \left\{\left(\tilde{x}_j, \tilde{y}_j\right)\right\}(j = 1, \ldots, n)$
(3) model $\mathcal{F}$
(4) loss function $\mathcal{L}$
(6) the number of iterations I
(7) physical transformation T

Output:
(1) parameters of model θ

Initialize θ
$s \leftarrow 0$
while $s < I$ **do**
  for $k = 0$ to n **do**
    Randomly choose 0 or 1 as $flag$
    if $flag = 1$ **then**
      $\tilde{x}_k \leftarrow T(\tilde{x}_k)$
    end if
  end for
  Update model parameters:
    $\theta_{s+1} \leftarrow \theta_s - \nabla_{\theta_s}\{\mathcal{L}(\mathcal{F}(\theta_s; x), y) + \mathcal{L}(\mathcal{F}(\theta_s; \tilde{x}), y_t)\}$
  $s \leftarrow s + 1$
end while
return θ

3.3. Physical transformations

The backdoor attacks are restricted by various physical conditions, which will significantly degrade the attack performance. Inspired by existing robust physical adversarial example attacks (Athalye et al., 2018; Chen et al., 2018; Eykholt et al., 2018), the proposed PTB method aims to model the possible physical constraints (that a backdoor trigger may experience in real physical world) in advance to ensure the robustness of the backdoor attack. Specifically, at each iteration of the model training, the proposed method performs physical transformations on each injected backdoor instance, i.e., $T(x + \delta)$, where T represents the five physical transformations, as shown in Fig. 3.

To guarantee that the attacker can also successfully launch the backdoor attack under normal physical conditions, at each iteration of the training, there is a 50% probability of not undergoing any transformations (keeps unchanged), as shown in Fig. 3. In other words, the proposed method transforms each backdoor face image with a certain probability (50% in this paper) at each iteration of

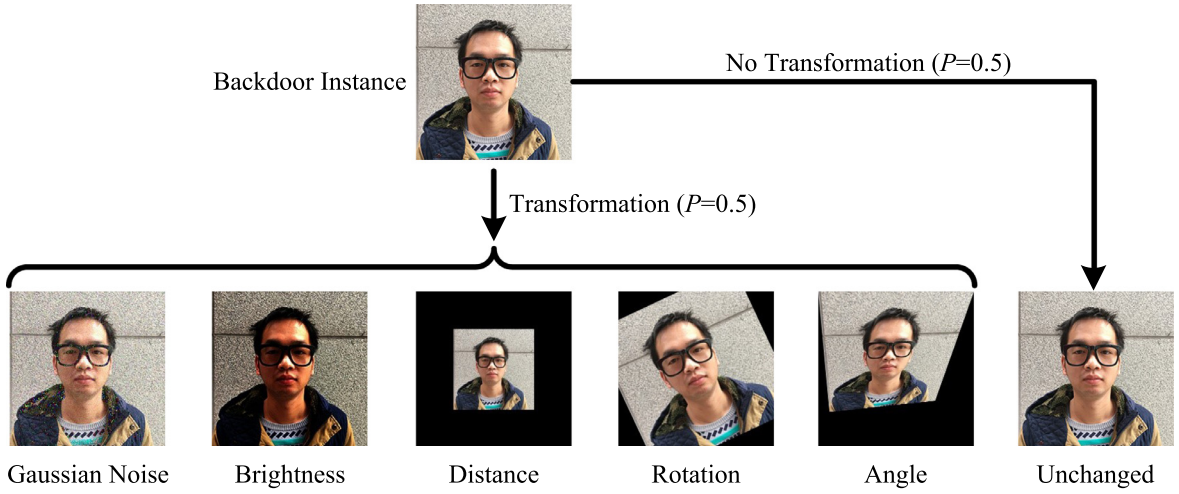


Fig. 3. The physical transformations performed on the injected backdoor face images at each iteration of model training.

training, which can be formulated as:

$$p \cdot T(x+\delta) + (1-p) \cdot (x+\delta), \quad p \in [0,1] \tag{2}$$

where p denotes the probability of transformation, which takes value of 0 or 1 with 50% probability.

In this way, for the proposed robust physical backdoor attack method, the objective function is formalized as:

$$\max_{p \in [0,1]} P_r\{F[p \cdot T(x+\delta) + (1-p) \cdot (x+\delta)] = y_t\} \tag{3}$$

In this paper, we consider five physical transformations that a backdoor trigger most likely to experience during the face recognition process in real physical world, i.e., T={Angle, Distance, Rotation, Brightness, Gaussian Noise}. The transformations of the proposed backdoor attack method at each iteration are presented in Fig. 3.

- **Angle.** The *Angle* transformation rotates the backdoor face image at a random angle (vertical and horizontal) that ranges from 0° to 90°. The horizontal rotation (i.e., left-right) simulates the physical constraints in which the camera captures face images at different horizontal angles, while the vertical rotation (i.e., up-down) simulates the physical constraints in which the face images are captured at different vertical angles.
- **Distance.** The *Distance* transformation simulates the distance changes between the attacker and the camera when the images are captured by the camera at different distances. To simulate the distance changes, the injected backdoor face images are scaled at random (ranging from 0.8 to 1.2 times). As a result, the backdoor attack can be successfully launched at different distances.
- **Rotation.** The *Rotation* transformation simulates the rotation of the backdoor instance that caused by the swing of an attacker's face. In each iteration, the backdoor face images are rotated by a random angle on 2D plane.
- **Brightness.** To make sure that the physical backdoor attack can be successfully launched under different lighting conditions, the proposed method performs the *Brightness* transformation to increase or decrease the brightness of the backdoor face images during training, in order to make the backdoor trigger can adapt to different lighting conditions.
- **Gaussian noise.** Since the backdoor face images are captured by a camera before being input into the face recognition model, the random environment noise introduced by the camera will affect the effectiveness of the backdoor trigger. The *Gaussian noise* transformation adds random gaussian noises on the injected backdoor face images.

3.4. Can existing countermeasures defend against physical backdoor attacks?

All existing backdoor defense works aim at detecting or mitigating the backdoor attacks in digital domain, where the backdoor attack is a fixed pattern. In the digital backdoor attack, in order to successfully trigger the backdoor, the shape, size, intensity and position of the trigger need to be the same as that of the trigger in the training stage. In comparison, in physical world, the trigger is usually a physical object (such as a pair of glasses), which needs to be captured by a camera from the physical world. Due to the various physical constraints introduced during the image capturing and processing, the captured physical trigger is totally different from the one used in the training stage. As a result, the physical trigger will fail to trigger the backdoor. However, through the proposed robust physical backdoor attack method, the enhanced physical trigger captured by the camera can still perform the backdoor attack even its shape, size and position has changed compared with the one used in the training stage. The existing countermeasures are based on the assumption that the trigger used to perform the attack is the same as the one injected into the training set. However, the proposed robust physical backdoor attack breaks this assumption, as the enhanced physical trigger captured by a camera is not the same as the one used in the training stage. Moreover, the physical triggers used in physical backdoor attack are real objects. In comparison, the digital triggers are some digital patterns. Since the captured physical trigger is more natural than the digital trigger which is directly patched onto the clean image, the features of physical objects are similar to those benign features that the model learned from clean images. Hence, the physical trigger is more difficult to detect compared with digital trigger. In summary, in *pixel space*, the physical trigger is different from the one used in the training stage, and in *feature space*, the physical trigger behaves like a benign feature. As a result, existing backdoor defense methods will perform poorly against the physical backdoor attacks.

To date, there is no defense method designed to detect the physical backdoor attacks. We choose three backdoor defense methods, STRIP (Gao et al., 2019), SentiNet (Chou et al., 2020) and Activation Clustering (Chen et al., 2019), to evaluate the robustness of the proposed physical backdoor attack method against existing defenses in Section 5.

4. Experimental results

4.1. Experimental setup

Dataset. In this work, the experimental evaluations are conducted on two large and realistic datasets, the YouTube Aligned Face dataset (Wolf et al., 2011), and the VGG Face dataset (Parkhi et al., 2015). The YouTube Aligned Face dataset (Wolf et al., 2011) contains the face images of 1595 different persons, and each person has tens to thousands face images. The VGG Face dataset (Parkhi et al., 2015) is a facial recognition dataset, which consists of 2,604,849 images from 2622 persons. In our experiments, all the face images are cropped and resized to 224 × 224. For each dataset, we randomly select 100 persons from the dataset as the experimental data, where each person has 120 face images (100 images for training and 20 images for testing). In this way, our experimental dataset contains a total of 12,000 face images.

DNN models. In the experiments, the target DNN model is VGGFace (Parkhi et al., 2015). VGGFace is a 16-layer face recognition model, which consists of 13 convolutional layers and 3 fully connected layers (Parkhi et al., 2015). We adopt the settings in exisiting works (Chen et al., 2017; Wenger et al., 2021) and use the VGGFace model that pre-trained on the ImageNet dataset (Deng et al., 2009). We fine-tune the last three layers of the VGGFace model, and replace the softmax activation layer, so as to train the VGGFace model on our experimental dataset. The model is fine-tuned on YouTube Aligned Face dataset and VGG Face dataset respectively, where the training epoch is set to 30, and the batch size is set to 64. In our experiment, the test accuracy of VGGFace model without the backdoor attacks is 96.33% and 92.04% on YouTube Aligned Face dataset and VGG Face dataset respectively.

Metrics. The proposed PTB method is evaluated with the following two metrics.

- Backdoor attack success rate SR_{bd} (Gu et al., 2019). This metric represents the proportion of backdoor face images that are classified as the target class y_t among all the submitted backdoor face images, which is calculated as follows:

$$SR_{bd} = \frac{N_t}{N_s} \times 100\% \tag{4}$$

where N_t represents the number of backdoor face images that classified as the target class y_t, and N_s represents the number of all the submitted backdoor face images.

- Performance drop of the DNN model A_{drop} (Xue et al., 2020a). This metric represents the degradation of the model's test accuracy caused by the proposed backdoor attack, which is calculated as follows:

$$A_{drop} = A_{cl} - A_{bd} \tag{5}$$

where A_{cl} represents the test accuracy of the model that trained on clean face images, and A_{bd} represents the test accuracy of the model that trained on backdoored training set.

Backdoor face images. In the experiment, we uses three physical objects as the physical trigger, respectively. The backdoor face images are captured by iPhone. All the backdoor face images are color images and are resized to 224 × 224.

- Three triggers. As shown in Fig. 4, the three different physical triggers are a 4 cm × 4 cm black square, a black triangle, and a pair of black glasses. In order to evaluate the performance of the physical backdoor attack when the trigger is on different positions of the attacker's face, we also capture the photos of the attacker with the square trigger on his chin (as shown in Fig. 4d).

- Simple scene & complex scene. In order to evaluate the performance of the proposed robust physical backdoor attack under different physical conditions, we take photos of the attacker injected with triggers under two attack scenarios, *Simple Scene* and *Complex Scene*. *Simple Scene* represents the ideal physical condition, in which the backdoor instances are not restricted by physical conditions during the attack. *Complex Scene* represents the complex physical scenario, in which the backdoor instances suffer from various complex physical conditions during the attack. These complex physical conditions are as follows: (i) distance; (ii) rotation; (iii) random noise; (iv) lighting condition; (vi) wide angle; (vii) elevation & depression angle.

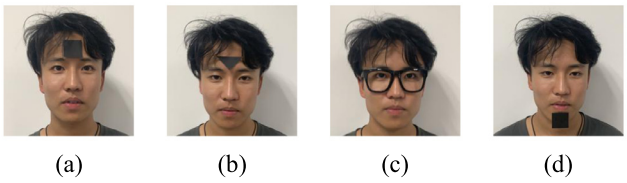


Fig. 4. Examples of backdoor instances injected with different triggers: (a) Square; (b) Triangle; (c) Glasses; (d) Square on chin.

Fig. 5. Examples of backdoor instances captured under different physical conditions.

Fig. 5 shows some backdoor face images captured under different complex physical conditions. As shown in Fig. 5: images in the first column are backdoor face images captured from a wide angle; images in the second column are backdoor face images captured from a elevation angle; images in the third column are backdoor face images captured from a depression angle; images in the last column are images captured under dim-lighting condition.

We capture the backdoor face images for each of the 3 triggers (square, triangle, glasses) under *Simple Scene* to generate backdoor instances in the training phase, and we capture the backdoor face images for each of the 4 triggers (square, triangle, glasses, square on chin) under each scene (*Simple Scene, Complex Scene*) to evaluate the performance of backdoor attack in the testing phase. In the training phase, we collect 50 backdoor face images for each trigger under *Simple Scene* to train the model on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), respectively. In the testing phase, we collect 20 and 50 backdoor face images for each trigger under each scene (*Simple Scene, Complex Scene*) to evaluate the model trained on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), respectively. To statistically evaluate the performance of the proposed method, for each trigger, the backdoor embedding and attacking process are repeated for 5 times to train

Table 2
The time overhead of training the model on clean training set and backdoor training set.

Dataset	YouTube Aligned Face		VGG Face	
	Clean	Backdoor	Clean	Backdoor
Training Time	3460s	3468s	3478s	3487s
Extra Time Overhead	8s		9s	

Table 3
Physical attack performance of backdoor trigger *square* on YouTube Aligned Face dataset and VGG Face dataset.

Dataset	Method	Simple Scene			Complex Scene		
		min	max	ave	min	max	ave
YouTube Aligned Face	without PTB	75%	100%	**91%**	0%	25%	**5%**
	with PTB	95%	100%	**99%**	65%	90%	**78%**
VGG Face	without PTB	60%	94%	**76%**	14%	26%	**21%**
	with PTB	86%	98%	**94%**	70%	88%	**78%**

5 different models on each dataset (YouTube Aligned Face dataset, VGG Face dataset), and each time we choose a different target victim. In practice, the attacker only needs to implement the attacking process once to launch the backdoor attack. For each trigger under each scene (*Simple, Complex*), we submit 20 and 50 photos to evaluate the 5 models trained on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), respectively. Hence, for each trigger under each scene on each dataset, there are 5 attack results, and we report the minimum, average, maximum among these 5 results.

4.2. Experimental results

We implement two types of backdoor attacks on the face recognition model. First, without the proposed PTB method, these backdoor face images are directly injected into the clean training set to train the face recognition model so as to embed the backdoor. Second, with the proposed PTB method, we perform a series of physical transformations on these injected backdoor face images during each iteration of the training process. In addition, the backdoor attacks are implemented in aforementioned two attack scenarios: *Simple Scene* and *Complex Scene*. Specifically, in the experiments, in order to evaluate the effectiveness of the proposed method, we collect the attackers' face images with the backdoor trigger under the *Simple* and *Complex* scenes, respectively.

4.2.1. Attack effectiveness

We evaluate the proposed method on the VGGFace model (Parkhi et al., 2015) that trained on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), respectively. We use a black square with the size of 4 cm $\times$ 4 cm as the trigger (referred as *Square*). In the real physical world, we paste the *Square* trigger to the forehead of the attacker, and use an iPhone to capture the photos under the *Simple* and *Complex* scenes, respectively. Then, these backdoor images are submitted to the face recognition model to evaluate the performance of the proposed backdoor attack.

Table 3 shows the performance of backdoor attacks on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015). Without the proposed PTB method, the performance of backdoor attacks under the *Simple* scene is high, where the maximum and average attack success rates are 100% (maximum) and 91% (average) on YouTube Aligned Face dataset, 94% (maximum) and 76% (average) on VGG Face dataset, respectively. However, the attack success rate without the proposed PTB method drops sharply under the *Complex* scene, where the attack success rates are only 0% (minimum), 25% (maximum), 5% (average) on YouTube Aligned Face dataset (Wolf et al., 2011), and 14% (minimum), 26% (maximum), 21% (average) on VGG Face dataset (Parkhi et al., 2015), respectively. With the proposed PTB method, under the *Simple* scene, the backdoor attack success rate is 99% (average), 100% (maximum), 95% (minimum) on YouTube Aligned Face dataset, and 94% (average), 98% (maximum), 86% (minimum) on VGG Face dataset, respectively. Meanwhile, under the *Complex* scene, the highest attack success rate is high up to 90% and 88% on YouTube Aligned Face dataset and VGG Face dataset, while the average attack success rate is 78% on both YouTube Aligned Face dataset and VGG Face dataset. The above experimental results indicate that the backdoor attacks without the proposed PTB method completely fail under complex physical conditions. In comparison, after using the PTB method, the average attack success rate of backdoor attacks under complex physical conditions has increased from 5% to 78% (YouTube Aligned Face dataset) and from 21% to 78% (VGG Face dataset), which demonstrate the effectiveness of the proposed method. In addition, under *Simple* conditions, the attack performance with the proposed PTB method is also better than that without the PTB method.

To evaluate the extra training time overhead introduced by embedding the backdoor, we report the time cost of training clean model and backdoored model, respectively. Since we transform n backdoor instances and inject them into the training set, processing and learning these extra n backdoor instances will cause extra training time overhead. The experiment is run on Ubuntu 20.04 LTS with AMD 2700x (CPU) and GTX 1080 (GPU). For YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), the training time on clean training set and backdoor training set are reported in Table 2. The extra time overhead caused by processing and learning backdoor instances during the training is 8 seconds and 9 seconds on YouTube Aligned Face dataset and VGG Face dataset, respectively. Compared with the time overhead of the whole training process (3460 seconds on YouTube Aligned Face dataset and 3478 seconds on VGG Face dataset), the introduced extra time overhead is negligible.

4.2.2. Parameters discussion

- Different triggers. The proposed method is effective and robust for different types of backdoor triggers. To demonstrate this, in addition to the *Square* trigger, we also evaluate the effectiveness of the proposed method with other types of triggers. Specifically, a black triangle (referred as *Triangle*) and a pair of black-frame glasses (referred as *Glasses*) are used as the physical trigger of backdoor attacks, respectively. Similarly, we paste these two triggers on the attackers' faces respectively and then take photos to perform the physical backdoor attacks. The attack performance of the two different triggers on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015) in real world are presented in Table 4. Table 4 shows that, without the PTB method, the backdoor attacks perform well under the *Simple* scene, where the average attack success rate is 96% (*Triangle*) and 96% (*Glasses*) on YouTube Aligned Face dataset (Wolf et al., 2011), 86% (*Triangle*) and 99% (*Glasses*) on VGG Face dataset (Parkhi et al., 2015), respectively. However, under the *Complex* scenes, the performance of these two backdoor attacks without the PTB method are rather poor. The average attack success rate is only 11% (*Triangle*) and 9% (*Glasses*) on YouTube Aligned Face dataset, 20% (*Triangle*) and 13% (*Glasses*) on VGG Face dataset, respectively. Once the proposed PTB method has been applied, the performance of these two backdoor attacks are greatly improved. The average attack success rate under the *Complex* scene is 82% (*Triangle*) and 79% (*Glasses*) on YouTube Aligned Face dataset (Wolf et al., 2011), 86% (*Triangle*) and 85% (*Glasses*) on VGG Face dataset (Parkhi et al., 2015), respectively. Meanwhile, under the simple scene, the attack per-

Table 4
Attack performance of two different backdoor triggers in the physical world on YouTube Aligned Face dataset and VGG Face dataset.

Dataset	Trigger	Method	Simple Scene			Complex Scene		
			min	max	ave	min	max	ave
YouTube Aligned Face	Triangle	without PTB	85%	100%	**96%**	0%	35%	**11%**
		with PTB	85%	100%	**97%**	60%	95%	**82%**
	Glasses	without PTB	90%	100%	**96%**	0%	20%	**9%**
		with PTB	100%	100%	**100%**	70%	90%	**79%**
VGG Face	Triangle	without PTB	76%	96%	**86%**	2%	38%	**20%**
		with PTB	92%	100%	**96%**	76%	98%	**86%**
	Glasses	without PTB	97%	100%	**99%**	2%	46%	**13%**
		with PTB	100%	100%	**100%**	62%	96%	**85%**

Table 5
Attack performance of the proposed PTB method on YouTube Aligned Face dataset and VGG Face dataset when the backdoor trigger is pasted on other positions (i.e., on the chin of an attacker's face).

Dataset	Method	Simple Scene			Complex Scene		
		min	max	ave	min	max	ave
YouTube Aligned Face	without PTB	85%	90%	**87%**	0%	30%	**10%**
	with PTB	90%	100%	**97%**	65%	85%	**75%**
VGG Face	without PTB	42%	92%	**69%**	26%	60%	**38%**
	with PTB	58%	94%	**78%**	54%	92%	**74%**

formance with the PTB method is also higher than that without the PTB method. Specifically, the average attack success rate with PTB under the *Simple* scene is 97% (*Triangle*) and 100% (*Glasses*) on YouTube Aligned Face dataset, 96% (*Triangle*) and 100% (*Glasses*) on VGG Face dataset, respectively.

- Different positions. In the above experiments, the triggers (*Square, Triangle* and *Glasses*) are pasted on the forehead of the attackers' faces. In fact, for a backdoor trigger that pasted on other positions, the proposed method is also effective and feasible. To demonstrate this, we paste a black square with the size of 4 cm $\times$ 4 cm on the chin of the attackers' faces to evaluate the impacts of triggers pasted on different positions. The experimental results are shown in Table 5. The average success rate without PTB under *Simple* scene is only 87% and 69% on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), respectively. In comparison, the average success rate with PTB under *Simple* scene is 97% and 78% on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), respectively. Under the *complex* scene, once the PTB method is exploited, the average attack success rate of *Square* trigger has improved from 10% (without PTB) to 75% (with PTB) on YouTube Aligned Face dataset, and has improved from 38% (without PTB) to 74% (with PTB) on VGG Face dataset. The above experimental results demonstrate that when the trigger is pasted on the chin of the attacker's face, the proposed method is still effective.

- Different injection numbers. Lastly, we evaluate the impacts of different numbers of injected backdoor face images on the proposed method. Specifically, we select the person with the label of "022" in the YouTube Aligned Face dataset (Wolf et al., 2011) and the person with the label of "017" in VGG Face dataset (Parkhi et al., 2015) as the target victim respectively, and use the backdoor trigger *Square* (YouTube Aligned Face), *Glasses* (VGG Face) to implement the backdoor attacks. We train 5 models with different number of injected backdoor instances (5, 10, 20, 50, 100) on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015), respectively. For each testing, we submit 20 backdoor face images with trigger *Square* (YouTube Aligned Face dataset) and 50 backdoor face images with trigger *Glasses* (VGG Face dataset) under each scene (*Simple, Complex*) to evaluate the backdoor attack success rate, respectively. The backdoor face images generation and the detailed attack process are the same as that described in Section 4.2.1.

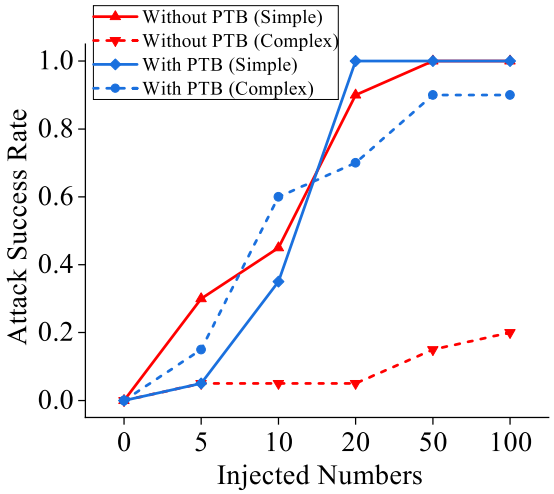


Fig. 6. The attack success rate under different numbers of injected backdoor instances on YouTube Aligned Face dataset.

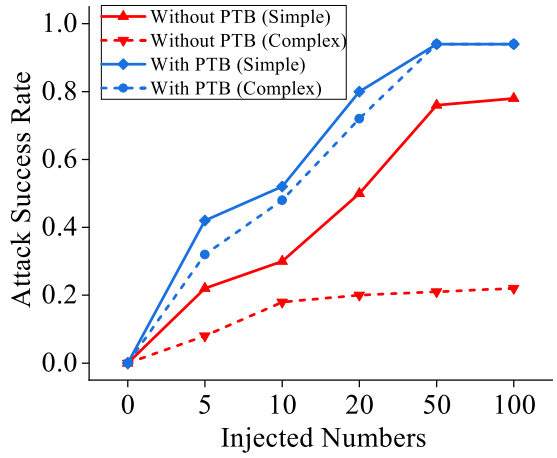


Fig. 7. The attack success rate under different numbers of injected backdoor instances on VGG Face dataset.

The impacts of the number of injected backdoor instances on the attack success rate of the proposed attack method on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015) are presented on Fig. 6 and Fig. 7, respectively. The results shows that: (i) As the injection number increases, the performance of the backdoor attacks with PTB under both simple scene and complex scene improves. Specifically, as the injected

number of backdoor instances increases from 0 to 100, the success rate of the proposed physical backdoor attack with PTB under *Complex Scene* increases from 0% to 90% (YouTube Aligned Face), and from 0% to 94% (VGG Face), respectively. Under *Simple Scene*, the success rate of backdoor attack with PTB also increases from 0% to 100% (YouTube Aligned Face) and from 0% to 94% (VGG Face), respectively. (ii) When the injection number reaches 50, which only accounts for 0.5% (50/10000) of the training set, the backdoor attack performance with PTB has already reached a high value. Specifically, when the number of injected backdoor instances is 50, the backdoor attack success rate with PTB reaches 100% (*Simple Scene*), 90% (*Complex Scene*) on YouTube Aligned Face dataset (Wolf et al., 2011) and 94% (*Simple Scene*), 94% (*Complex Scene*) on VGG Face dataset (Parkhi et al., 2015). (iii) Under the *Complex* physical conditions, the proposed PTB method significantly increases the backdoor attack success rate. Specifically, when injecting 50 backdoor instances, the attack success rate without PTB is only 15% (YouTube Aligned Face dataset), 21% (VGG Face dataset), while the attack success rate with PTB is 90% (YouTube Aligned Face dataset), and 94% (VGG Face dataset), respectively. (iv) Generally, under the *Simple* physical conditions, the proposed PTB method also improves the backdoor attack success rate.

5. Robustness against existing backdoor defenses

In this section, we evaluate the robustness of the proposed physical backdoor attack against three state-of-the-art backdoor defenses, STRIP (Gao et al., 2019), SentiNet (Chou et al., 2020) and Activation Clustering (Chen et al., 2019).

5.1. Robustness of the proposed method against STRIP

STRIP (Gao et al., 2019) aims to detect whether an input image contains a trigger at the inference stage. STRIP intentionally adds a set of clean images to the input image in order to generate a set of blended images. Then, STRIP calculates the entropy of the predicted labels of these blended images (Gao et al., 2019). The entropy of a backdoor instance is significantly lower than the entropy of a clean image. A low entropy indicates that the corresponding input image is a backdoor instance (Gao et al., 2019).

In this experiment, for each backdoor instance, a set of clean images are randomly selected to superimpose onto the backdoor instance, so as to generate a set of blended backdoor instances (Gao et al., 2019). Then the entropy is calculated from the predictions of the model on these blended backdoor instances (Gao et al., 2019). For YouTube Aligned Face dataset, for each trigger, we use 20 backdoor face images captured under *Simple* scene and 20 clean images selected from YouTube Aligned Face dataset (Wolf et al., 2011) to calculate the entropy distribution. For VGG Face dataset, for each trigger, we use 50 backdoor face images captured under *Simple* scene and 50 clean images selected from VGG Face dataset (Parkhi et al., 2015) to calculate the entropy distribution. The entropy distribution of backdoor face images and clean images on the two datasets is shown in Fig. 8. STRIP (Gao et al., 2019) aims to find a proper entropy threshold to distinguish the clean images from the backdoor instances. However, as shown in Fig. 8, for the proposed robust physical backdoor attack method, the entropy distribution of the backdoor instances is similar with that of clean images. Specifically, the entropy distribution of the blended clean images is mostly overlapped with that of the blended backdoor instances. Hence, STRIP (Gao et al., 2019) cannot find an entropy threshold to distinguish between the clean images and backdoor instances. As a result, STRIP (Gao et al., 2019) fails to detect the physical backdoor face images.

Generally, in digital domain, STRIP (Gao et al., 2019) can achieve a high detection rate on digital triggers which have high intensity, as the high-intensity trigger is still visible and salient even after being blended with other images. However, due to the constraints of various physical conditions, the intensity of physical trigger captured by a camera is relatively low. Therefore, after the blending process of STRIP, the physical trigger is easy to be destroyed by the clean images and become ineffective. The examples of digital trigger and physical trigger are shown in Fig. 9. As shown in the first row of Fig. 9, the digital trigger remains visible and salient even after being blended with a clean image. However, for physical backdoor, as shown in the second row of Fig. 9, after the blending process, the physical trigger is destroyed by the clean image and thus cannot activate the backdoor. Therefore, the predictions of blended backdoor instances will be random, just like that of blended clean images. In this way, the entropy of predictions of blended backdoor instances are similar to that of blended clean images. Thus, STRIP cannot find an entropy threshold to distinguish between the clean images and backdoor instances. As a result, in the experiment, STRIP (Gao et al., 2019) fails to detect the proposed physical backdoor attack.

5.2. Robustness of the proposed method against SentiNet

Given an input image, SentiNet (Chou et al., 2020) utilizes Gradient-weighted Class Activation Map (Grad-CAM (Selvaraju et al., 2017)) to locate a continuous region which contributes heavily to the predicted result. Then, SentiNet (Chou et al., 2020) carves out the region and patches it onto a set of clean images. If most of these clean images are predicted as the same label that is different from their ground-truth labels, the region is considered to contain a trigger (Chou et al., 2020). Correspondingly, the input image is considered to be a backdoor instance.

The output of Grad-CAM (Selvaraju et al., 2017) is a saliency map of the input image. The most salient region of the input image can be located based on this saliency map. Because the SentiNet (Chou et al., 2020) method mainly depends on whether Grad-CAM (Selvaraju et al., 2017) can locate the correct region containing the trigger, we utilize Grad-CAM (Selvaraju et al., 2017) to generate the saliency map from the backdoor face images. If the most salient region does not include the trigger, SentiNet (Chou et al., 2020) will be considered to fail to detect the backdoor. In this experiment, for each physical backdoor trigger under *Simple* and *Complex* scenes, the saliency map is generated from the corresponding backdoor face image through Grad-CAM. The output of Grad-CAM only roughly highlights the location of the salient regions, which lacks fine details (e.g., shape and texture) in the highlighted regions (Selvaraju et al., 2017). Therefore, in the work (Selvaraju et al., 2017), Guided Backpropagation is combined with Grad-CAM to show fine-grained details, which is referred to as Guided Grad-CAM (Selvaraju et al., 2017). Note that, the Guided Grad-CAM does not change the salient regions that Grad-CAM locates, and it presents the fine-grained details of the salient region.

The output of Grad-CAM and Guided Grad-CAM generated from backdoor face images on the model trained on YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015) are shown in Fig. 10 and Fig. 11, respectively. In Fig. 10 and Fig. 11, the first column is the original backdoor face images. The second column is the output of Grad-CAM on the original backdoor face images. The third column is the output of Guided Grad-CAM corresponding to the images in the second column. Guided Grad-CAM shows the fine-grained salient features located by Grad-CAM. As shown in Fig. 10 and Fig. 11, Grad-CAM (Chou et al., 2020) incorrectly locates the background objects or other facial features as the potential backdoor region. The experimental results show that, for different physical triggers in backdoor face images captured under different scenes, Grad-CAM fails to locate the correct region containing physical backdoor trigger, which means SentiNet

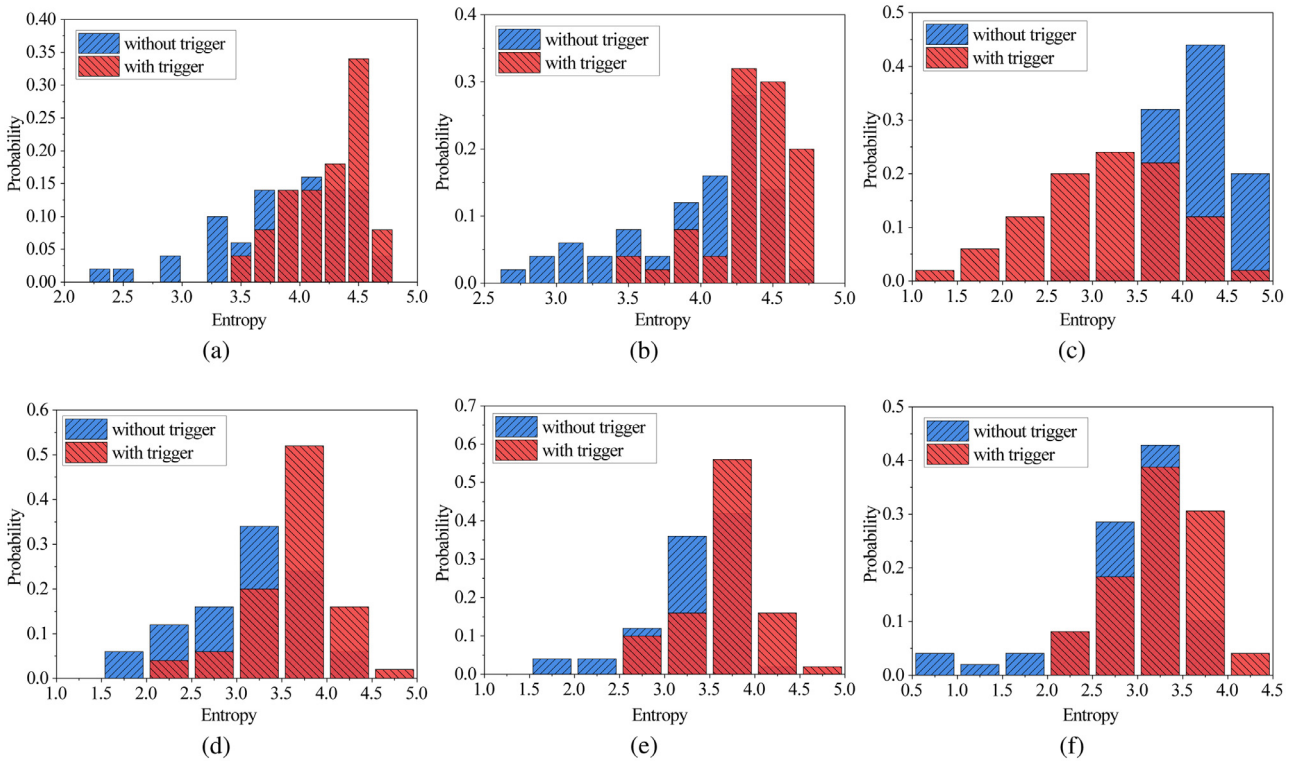


Fig. 8. The entropy distribution of the backdoor instances and clean images on two datasets. The first row is the results on the YouTube Aligned Face dataset (Wolf et al., 2011) and the second row is the results on VGG Face dataset (Parkhi et al., 2015). The trigger in the backdoor instances is: (a) Square; (b) Triangle; (c) Glasses; (d) Square; (e) Triangle; (f) Glasses.

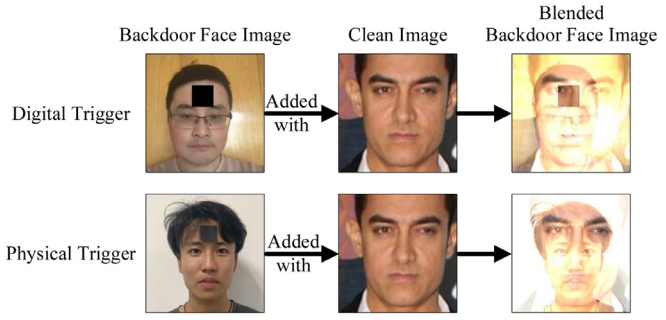


Fig. 9. Examples of blended backdoor face images generated from digital backdoor face images and physical backdoor face images.

(Chou et al., 2020) fails to detect the proposed physical backdoor attack.

In digital domain, SentiNet (Chou et al., 2020) is an effective defense method against digital backdoor attacks. However, the experimental results show that SentiNet (Chou et al., 2020) is ineffective against the proposed robust physical backdoor attack. The reasons are as follows: (i) The images in digital dataset are all pre-processed. Take the face recognition dataset as an example, the region of human face in the digital image is identified and the other content of image which is irrelevant to the face region is removed. Therefore, all face images in the dataset have no background or little background. However, in physical domain, the backdoor face images are captured from the physical world, so the captured backdoor face images inevitably have complicated backgrounds. As shown in Fig. 12, the left face image is an example from VGG Face dataset (Parkhi et al., 2015). The whole image apparently has little background and the human face occupies most space of the image. In comparison, the right face image is a photo taken in real physical world. The captured face image has a complicated background that contains various irrelevant objects. As a result, Grad-CAM will be influenced by the objects in the background and fails to locate the trigger on the human face, causing SentiNet (Chou et al., 2020) to fail to detect the proposed robust physical backdoor. (ii) As discussed in Section 5.1, the captured physical trigger has experienced several physical constraints, so it is less salient than the digital trigger which is directly added to the image. Meanwhile, in a face recognition task, except the physical trigger, there are many facial features (such as a big nose or blue eyes) which are also quite salient. These salient facial features are also the important factors which contribute heavily to the model's final prediction. Grad-CAM (Selvaraju et al., 2017) is likely to locate these facial features as the most salient region, as it aims to locate a region which contributes heavily to the prediction result. In this way, the most salient region located by Grad-CAM contains some benign facial features rather than the physical backdoor. Therefore, SentiNet (Chou et al., 2020) fails to detect the proposed physical backdoor attack.

5.3. Robustness of the proposed method against Activation Clustering

Activation Clustering (AC) (Chen et al., 2019) aims to detect whether the training data contains backdoor instances. First, for each class, AC submits the training data of this class to the model and analyzes the activations of the last hidden layer. Second, AC clusters the training data of this class into two groups by applying 2-means clustering algorithm. Third, AC calculates one *Silhouette Score* according to the two clusters of training data of this class. A high *Silhouette Score* means the training data of this class has been backdoored. In this experiment, according to the experiment settings in AC (Chen et al., 2019), we set the threshold T_{ac} to 0.15. If the *Silhouette Score* is higher than 0.15, the training data of the

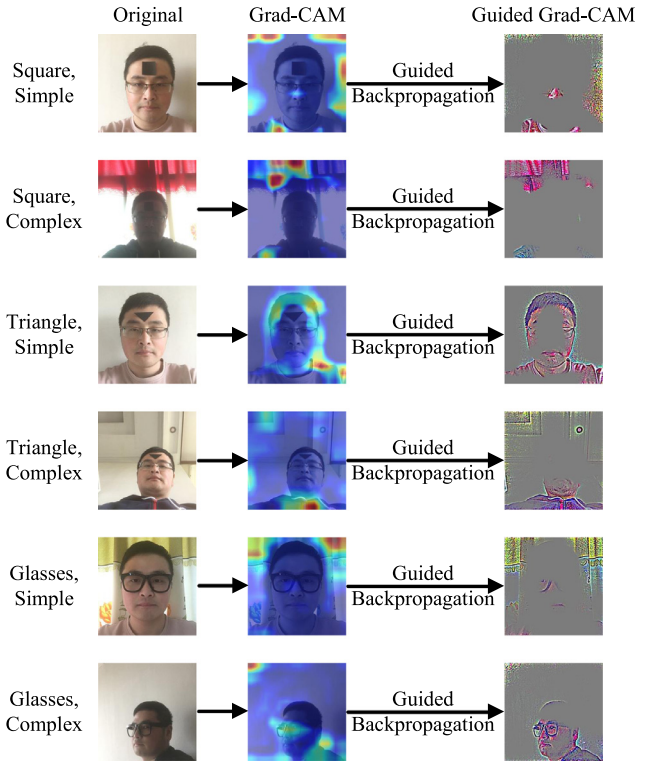


Fig. 10. The outputs of Grad-CAM and Guided Grad-CAM on backdoor face images with the model trained on YouTube Aligned Face dataset.

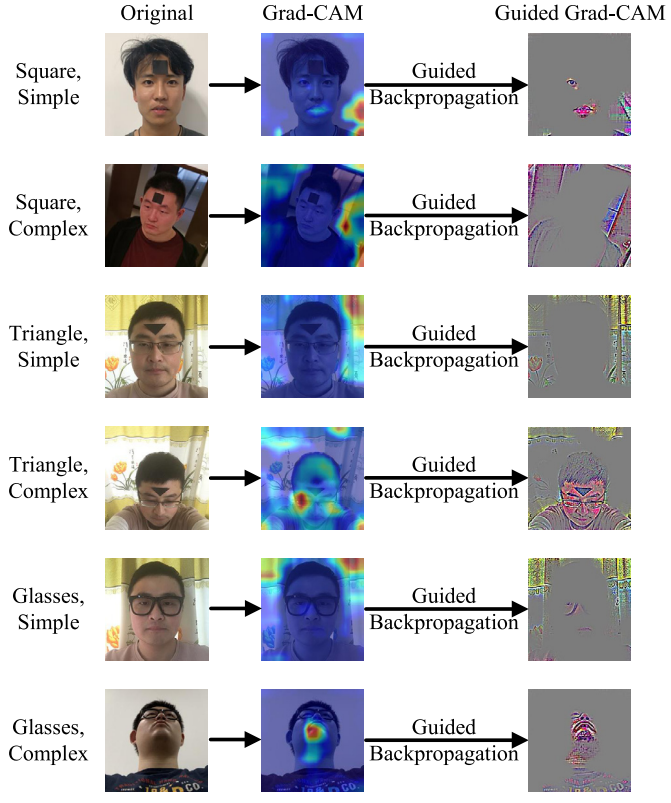


Fig. 11. The outputs of Grad-CAM and Guided Grad-CAM on backdoor face images with the model trained on VGG Face dataset.

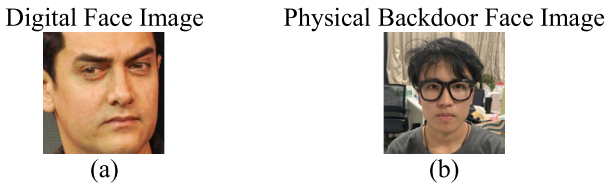


Fig. 12. Examples of digital face images and physical backdoor face images: (a) example of face images from VGG Face dataset; (b) example of physical backdoor face image.

corresponding class is considered to be backdoored. Otherwise, the training data of the corresponding class is considered to be clean.

Table 6 shows the detection results of Activation Clustering (Chen et al., 2019) against the proposed robust physical backdoor attack. The experimental results show that, for three backdoors (*Square, Triangle* and *Glasses* triggers) on two datasets, the *Silhouette Score* for target class are all lower than the threshold T_{ac}. Hence, Activation Clustering method fails to detect the proposed physical backdoor. The reason is that, AC distinguishes the backdoor instances from the clean images based on the assumption that backdoor instances active a different set of neurons compared with the clean images. However, the physical backdoor triggers used in this paper are real physical objects. Some features of these physical triggers are the same as that of clean facial images. Hence, the neurons activated by physical triggers will also be activated by clean images. In other words, the activations of the last hidden layer for physical triggers are similar with that for clean images. As a result, the Activation Clustering method cannot detect the proposed physical backdoor based on the activations of the last hidden layer.

In summary, all the three state-of-the-art defenses (Chen et al., 2019; Chou et al., 2020; Gao et al., 2019) which are effective against digital backdoor attacks fail to detect the proposed robust physical backdoor attack in real physical world.

6. Discussion

As discussed in the above sections, under strong attacker assumption, the attacker has access to both the training data and the training process of the target model (Cheng et al., 2021; Nguyen and Tran, 2020; 2021; Souri et al., 2021). The attacker injects backdoor instances into the training set and performs physical transformations on the backdoor instances during each iteration of training. In addition, the proposed robust physical backdoor attack method can also be applied under weak attacker scenario, where the attacker can only control a small portion of the training data and has no access to the training process, parameters or architecture of the target model, i.e., under black-box attack scenario. For the weak attacker scenario, the attacker can generate the transformed backdoor instances through the proposed PTB method in advance, and injects both transformed backdoor instances and unchanged backdoor instances into the training set. Then, the black-box model trained on the backdoored training set will be embedded with the backdoor. Hence, the weak attacker can also embed the backdoor into the target model without the control over the training process of the target model.

7. Conclusion

This paper proposes a robust physical backdoor attack method. By performing various physical transformations during each iteration of the training, the proposed backdoor attack method ensures the physical robustness and effectiveness of backdoor attacks. The physical transformations simulate these physical constraints that a

Table 6
The detection results of Activation Clustering method against the proposed robust physical backdoor attack.

Dataset	Backdoor Trigger	Target Label	Silhouette Score	Detection Result
YouTube	Square	23	0.0351	Clean
Aligned	Triangle	37	0.0368	Clean
Face	Glasses	22	0.0313	Clean
VGG	Square	0	0.0380	Clean
Face	Triangle	94	0.0269	Clean
	Glasses	41	0.0563	Clean

backdoor trigger may experience in real world, which can significantly improve its robustness under complex physical conditions. Experimental results on two face recognition datasets (YouTube Aligned Face dataset (Wolf et al., 2011) and VGG Face dataset (Parkhi et al., 2015)) demonstrate that, the proposed method can significantly improve the backdoor attack success rate on the modern face recognition model (VGGFace (Parkhi et al., 2015)), especially under the complex physical scenarios. Meantime, the normal performance of the models on clean inputs are not affected, thus the proposed backdoor attack is convert and is difficult to notice. Besides, experimental results also show that the state-of-the-art countermeasures designed for digital backdoor attacks fail to detect the proposed robust physical backdoor attack. In the future work, we will study the defenses against physical backdoor attacks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Mingfu Xue: Conceptualization, Methodology, Writing – review & editing. **Can He:** Software, Validation, Writing – original draft. **Yinghao Wu:** Software, Investigation, Validation, Writing – original draft. **Shichang Sun:** Software, Investigation, Validation, Writing – original draft. **Yushu Zhang:** Conceptualization. **Jian Wang:** Supervision. **Weiqiang Liu:** Conceptualization, Supervision.

Acknowledgment

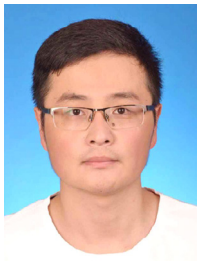
This work is supported by the National Natural Science Foundation of China (No. 61602241), and CCF-NSFOCUS Kun-Peng Scientific Research Fund (No. CCF-NSFOCUS 2021012).

References

Ali, M.A., Azad, M.A., Centeno, M.P., Hao, F., van Moorsel, A., 2019. Consumer-facing technology fraud: economics, attack methods and potential solutions. Future Gener. Comput. Syst. 100, 408–427.
Athalye, A., Engstrom, L., Ilyas, A., Kwok, K., 2018. Synthesizing robust adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning, pp. 284–293.
Bagdasaryan, E., Shmatikov, V., 2021. Blind backdoors in deep learning models. In: 30th USENIX Security Symposium, pp. 1505–1521.
Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V., 2020. How to backdoor federated learning. In: the 23rd International Conference on Artificial Intelligence and Statistics, pp. 2938–2948.
Bojarski, M., Testa, D.D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K., 2016. End to end learning for self-driving cars. arXiv:1604.07316.
Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I.M., Srivastava, B., 2019. Detecting backdoor attacks on deep neural networks by activation clustering. In: AAAI Workshop on Artificial Intelligence Safety, pp. 1–8.
Chen, S., Cornelius, C., Martin, J., Chau, D.H.P., 2018. ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector. In: Proceedings of Machine Learning and Knowledge Discovery in Databases, pp. 52–68.
Chen, X., Liu, C., Li, B., Lu, K., Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv:1712.05526.
Cheng, S., Liu, Y., Ma, S., Zhang, X., 2021. Deep feature space trojan attack of neural networks by controlled detoxification. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, pp. 1148–1156.
Chou, E., Tramèr, F., Pellegrino, G., 2020. SentiNet: Detecting localized universal attacks against deep learning systems. In: IEEE Security and Privacy Workshops, pp. 48–54.
Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. ImageNet: A large-scale hierarchical image database. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 248–255.
Désidéri, J.-A., 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. C.R. Math. 350 (5–6), 313–318.
Doan, K., Lao, Y., Zhao, W., Li, P., 2021. LIRA: Learnable, imperceptible and robust backdoor attacks. In: the IEEE/CVF International Conference on Computer Vision, pp. 11966–11976.
Dumford, J., Scheirer, W.J., 2020. Backdooring convolutional neural networks via targeted weight perturbations. In: IEEE International Joint Conference on Biometrics, pp. 1–9.
Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D., 2018. Robust physical-world attacks on deep learning visual classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625–1634.
Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S., 2019. STRIP: A defence against trojan attacks on deep neural networks. In: Proceedings of the 35th Annual Computer Security Applications Conference, pp. 113–125.
Garg, S., Kumar, A., Goel, V., Liang, Y., 2020. Can adversarial weight perturbations inject neural backdoors. In: The 29th ACM International Conference on Information and Knowledge Management, pp. 2029–2032.
Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S., 2019. BadNets: evaluating backdooring attacks on deep neural networks. IEEE Access 7, 47230–47244.
Guo, C., Wu, R., Weinberger, K.Q., 2020. TrojanNet: embedding hidden trojan horse models in neural networks. arXiv:2002.10078.
Hampel, F.R., 1974. The influence curve and its role in robust estimation. J Am Stat Assoc 69 (346), 383–393.
Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S.Z., Hospedales, T., 2015. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In: Proceedings of the IEEE international conference on computer vision workshops, pp. 142–150.
Li, S., Xue, M., Zhao, B., Zhu, H., Zhang, X., 2020. Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Trans Dependable Secure Comput 18 (5), 2088–2105. 1–1.
Li, Y., Zhai, T., Jiang, Y., Li, Z., Xia, S., 2021. Backdoor attack in the physical world. In: ICLR workshop on Robust and Reliable Machine Learning in the Real World, pp. 1–6.
Liu, K., Dolan-Gavitt, B., Garg, S., 2018. Fine-Pruning: Defending against backdooring attacks on deep neural networks. In: Proceedings of 21st International Symposium on Attacks, Intrusions, and Defenses, pp. 273–294.
Liu, Y., Ma, S., Aafer, Y., Lee, W., Zhai, J., Wang, W., Zhang, X., 2018. Trojaning attack on neural networks. In: Proceedings of the 25th Annual Network and Distributed System Security Symposium, pp. 1–17.
Liu, Y., Ma, X., Bailey, J., Lu, F., 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In: Proceedings of the 16th European Conference on Computer Vision, pp. 182–199.
Mejia, N., 2020. Facial recognition in banking current applications. https://emerj.com/ai-sector-overviews/facial-recognition-in-banking-current-applications/.
Nguyen, T.A., Tran, A., 2020. Input-aware dynamic backdoor attack. In: Advances in Neural Information Processing Systems 33, pp. 1–5.
Nguyen, T.A., Tran, A.T., 2021. WaNet – Imperceptible warping-based backdoor attack. In: 9th International Conference on Learning Representations, pp. 1–16.
Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition. In: Proceedings of the British Machine Vision Conference, pp. 1–12.
Pasquini, C., Böhme, R., 2020. Trembling triggers: exploring the sensitivity of backdoors in DNN-based face recognition. EURASIP J. Inf. Secur. 2020, 1–12.
Rakin, A.S., He, Z., Fan, D., 2020. TBT: targeted neural network attack with bit trojan. In: Conference on Computer Vision and Pattern Recognition, pp. 13195–13204.
Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788.
Ren, S., He, K., Girshick, R.B., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

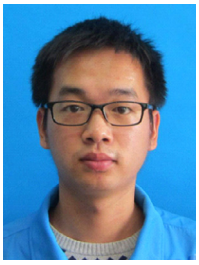
Saha, A., Subramanya, A., Pirsiavash, H., 2020. Hidden trigger backdoor attacks. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, pp. 11957–11965.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad–CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision, pp. 618–626.
Souri, H., Goldblum, M., Fowl, L., Chellappa, R., Goldstein, T., 2021. Sleeper agent: scalable hidden trigger backdoors for neural networks trained from scratch. arXiv:2106.08970.
Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y., 2019. Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks. In: IEEE Symposium on Security and Privacy, pp. 707–723.
Wenger, E., Passananti, J., Bhagoji, A.N., Yao, Y., Zheng, H., Zhao, B.Y., 2021. Backdoor attacks against deep learning systems in the physical world. In: Conference on Computer Vision and Pattern Recognition, pp. 6206–6215.
Wolf, L., Hassner, T., Maoz, I., 2011. Face recognition in unconstrained videos with matched background similarity. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 529–534.
Xue, M., He, C., Sun, S., Wang, J., Liu, W., 2021. Robust backdoor attacks against deep neural networks in real physical world. In: The 20th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 1–7.
Xue, M., He, C., Wang, J., Liu, W., 2020. One-to-N & N-to-One: two advanced backdoor attacks against deep learning models. IEEE Trans Dependable Secure Comput 1–17,early access. doi:10.1109/TDSC.2020.3028448.
Xue, M., He, C., Wang, J., Liu, W., 2021. Backdoors hidden in facial features: a novel invisible backdoor attack against face recognition systems. Peer-to-Peer Networking and Applications 14 (3), 1458–1474.
Xue, M., Yuan, C., Wu, H., Zhang, Y., Liu, W., 2020. Machine learning security: threats, countermeasures, and evaluations. IEEE Access 8, 74720–74742.
Yao, Y., Li, H., Zheng, H., Zhao, B.Y., 2019. Latent backdoor attacks on deep neural networks. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 2041–2055.
Zhang, Q., Ding, Y., Tian, Y., Guo, J., Yuan, M., Jiang, Y., 2021. Advdoor: Adversarial backdoor attack of deep learning system. In: 30th ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 127–138.
Zhong, H., Liao, C., Squicciarini, A.C., Zhu, S., Miller, D.J., 2020. Backdoor embedding in convolutional neural network models via invisible perturbation. In: Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, pp. 97–108.
Zou, M., Shi, Y., Wang, C., Li, F., Song, W., Wang, Y., 2018. PoTrojan: powerful neural-level trojan designs in deep learning models. arXiv: 1802.03043.

Mingfu Xue is currently an Associate Professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He received the Ph.D. degree from Southeast University, Nanjing, China, in 2014. From July 2011 to July 2012, he is a visiting Ph.D student in Nanyang Technological University, Singapore. He has been a technical program committee member for over 20 international conferences. He is a committee member of the Chinese artificial intelligence and security professional committee, a committee member of the Intelligent and Security Committee of Jiangsu Artificial Intelligence Society, and also an executive committee member of ACM Nanjing Chapter. He has been the Principal Investigator of 11 research projects and participated in 4 other research projects. He won the best paper award in ICCCS2015. His research interests include artificial intelligence security, secure and private machine learning systems, and hardware security.



Can He received the B.S. degree in computer science and technology from Anhui University of Technology, in 2017. He is currently pursuing the master's degree in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include artificial intelligence security, secure and private machine learning systems.

Yinghao Wu received the B.S degree in electronic science and technology from Nanjing University of Aeronautics and Astronautics in 2018. He is currently pursuing the master's degree in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include artificial intelligence security, secure and private machine learning systems.



Shichang Sun received the B.S. degree in computer science and technology from Nantong University, in 2019. He is currently pursuing the master's degree in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include artificial intelligence security, secure and private machine learning systems.

Yushu Zhang received the Ph.D. Degree from the College of Computer Science, Chongqing University, Chongqing, China, in Dec. 2014. He held various research positions at the City University of Hong Kong, Southwest University, University of Macau, and Deakin University. He is now a full Professor with College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. His research interests include multimedia security, artificial intelligence, cloud computing security, big data security, IoT security, and blockchain. He has published over 100 refereed journal articles and conference papers in these areas. He is an Editor of Signal Processing.

Jian Wang received the Ph.D. degree in Computer Application Technology from Nanjing University, China, in 1998. From 2001 to 2003, he was a postdoctoral researcher at the University of Tokyo, Japan. From 1998 to 2004, he was an associate professor in Nanjing University, China. He is currently a full professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China, where he was also the Vice Director of this college from 2010 to 2015. He is a committee member of the Chinese Cryptography Society, and was the Director of Jiangsu Provincial Cryptography Society. a His research interests include applied cryptography, system security, key management, security protocol, and information security. He has published over 70 papers in security related journals and international conferences.

Weiqiang Liu received the B.Sc. degree in Information Engineering from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China and the Ph.D. degree in Electronic Engineering from the Queen's University Belfast (QUB), Belfast, UK, in 2006 and 2012, respectively. In Dec. 2013, he joined the College of Electronic and Information Engineering, NUAA, where he is currently a Professor and the Vice Dean of the college. He has published one research book by Artech House and over 130 leading journal and conference papers. His paper was selected as the Highlight Paper of IEEE TCAS-I in the 2021 January Issue and the Feature Paper of IEEE TC in the 2017 December issue. He has been awarded the prestigious Excellent Young Scholar Award by National Natural Science Foundation of China in 2020. He serves as the Associate Editors for IEEE Transactions on Circuits and System I: Regular Papers (2020.1–2021.12), IEEE Transactions on Emerging Topics in Computing (2019.5–2021.4) and IEEE Transactions on Computers (2015.5–2019.4), an Steering Committee Member of IEEE Transactions on VLSI Systems (2021.1–2022.12). He is the program co-chair of IEEE ARITH 2020, and also technical program committee members for ARITH, DATE, ASAP, ISCAS, ASP-DAC, ISVLSI, GLSVLSI, SiPS, NANOARCH, AICAS and ICONIP. He is a member of CASCOM and VSA Technical Committee of IEEE Circuits and Systems Society. His research interests include approximate computing, hardware security and VLSI design for digital signal processing and cryptography.