IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 15, 2021, accepted August 30, 2021, date of publication September 3, 2021, date of current version September 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110239

Use Procedural Noise to Achieve Backdoor Attack

XUAN CHEN, YUENA MA, AND SHIWEI LU
Department of Base, Air Force Engineering University, Xi'an 710038, China

Corresponding author: Xuan Chen (chenxuan0839@gmail.com)

This work was supported by the Natural Science Foundation of Shaanxi Province of China under Grant 2021JM-216 and Grant 2021JQ-335.

ABSTRACT In recent years, more researchers pay their attention to the security of artificial intelligence. The backdoor attack is one of the threats and has a powerful, stealthy attack ability. There exist a growing trend towards the triggers is that become dynamic and global. In this paper, we propose a novel global backdoor trigger that is generated by procedural noise. Compared with most triggers, ours are much stealthy and straightforward to implement. In fact, there exist three types of procedural noise, and we evaluate the attack ability of triggers generated by them on the different classification datasets, including CIFAR-10, GTSRB, CelebA, and ImageNet12. The experiment results show that our attack approach can bypass most defense approaches, even for the inspections of humans. We only need poison 5%-10% training data, but the attack success rate(ASR) can reach over 99%. To test the robustness of the backdoor model against the corruption methods that in practice, we introduce 17 corruption methods and compute the accuracy, ASR of the backdoor model with them. The facts show that our backdoor model has strong robustness for most corruption methods, which means it can be applied in reality. Our code is available at https://github.com/928082786/pnoiseattack.

INDEX TERMS Deep learning, backdoor attack, procedural noise, global trigger.

I. INTRODUCTION

In recent years, deep learning has achieved great success in computer vision ([1]), natural language processing([2]) and graph neural network([3]). Besides, deep learning also has greatly promoted the epidemiological study of COVID-2019 and the development of antibodies.([4]–[6]). However, there exist many threats for deep learning: adversarial examples([7], [8]), which attack the models at the testing stage through adding some small perturbations that are unseen by humans; backdoor attack([9], [10]), which attack the models at training stage through inserting some poisoned data into a training dataset; inference attack([11], [12]), which infers the training data or the model weights according to the deep learning model could remember the training data. Therefore, we can see that deep learning is facing many security problems. The scenarios which the adversaries can attack includes face recognition([13]), disease diagnosis([14]), spam detection([15]), autonomous driving([16]) and so on. How to defend against these attacks is still an open problem for researchers. Compared with some attacks, the backdoor attack is stealthy, whose triggers are often not easily detected. However, the attacking ability of which is powerful, with a high attack success rate. Therefore, the backdoor attack is full of several research values. We assume that some individuals or small companies need to customize models to work with actual needs for the backdoor attack. However, these individuals or small companies have no enough resources to train models, so that they need to submit the requirements to the third-party platform, and the third-party platform will return the models to users after training. With such a process, the adversary can insert triggers into the model and have a backdoor model. The backdoor model can keep a high accuracy on benign inputs but classify the poison images as targeted labels that the adversary wants. Another scenario for the backdoor attack is that some companies or individuals will download datasets or backdoor models from the website. The datasets may be injected with the poisoned data, and the modes will be implanted with a backdoor if training model with these datasets. Even some large companies which have enough computing resources can also be attacked. Federated learning([17]) is a promising field that collects data from the million(even billion) devices in a real sense and trains models with these data to enhance the generalization ability of the model. However, the data collected from the reality is easy to be inserted into triggers and brings in backdoor attacks([18], [19])

There are a lot of attack methods in backdoor attacks, which can be classified as visible backdoor triggers([9], [10]) and invisible backdoor triggers([24], [25]). Except for image

The associate editor coordinating the review of this manuscript and approving it for publication was Nadeem Iqbal.

classification, some traces of backdoor attacks can also be found in these fields: video recognition([25]), natural language processing([26]), graph neural networks([27]). Although the backdoor attack has a powerful attack ability, the limitations of the triggers will affect the stealthiness of which. There exist many attack methods whose triggers can be detected by human vision. As we can see from Fig.1, the triggers of BadNets([9]), Blended([20]), SIG([21]), Refool([22]) are too apparent for the manual detection. Therefore, designing a more stealthy attack algorithm is full of necessary. In this paper, we introduce a novel trigger which generated by procedural noise. Fig.1 shows the poisoned examples generating by our approach, and we can find that which is more unnoticeable than any other poisoned examples. Moreover, our poisoned image is natural and difficult for humans to distinguish.

We introduce the related works in Section 2, including the works of attacks and defenses. In Section 3, we mainly propose the preliminary knowledge about procedural noise and backdoor attacks. Moreover, we will introduce our approach in Section 4, and the experiments for evaluating are in Section 5. In Section 6, we draw a conclusion of our paper.

The main contributions of our paper are following:

- We propose a novel backdoor attack approach, which is more powerful and stealthy. The triggers of our attack approach are easy to generate, which means that we can quickly generate lots of triggers, not static but still have attack ability.
- We introduce six different baseline defense approaches and evaluate our attack performance with these defenses on different datasets. The results prove that our approach is powerful and robust enough to defend most defense methods.
- To verify the robustness of our approach for the corruption methods, we list 17 corruption methods in the classification tasks and test the accuracy and ASR under these corruption methods. Our approach is still robust for the corruption methods, which means our approach has excellent application values.

II. RELATED WORK

In this section, we will introduce the related works on backdoor attacks and the defense approaches.

A. ATTACK

For backdoor attack, [9] firstly proposed BadNets attack. The BadNets attack only needs to insert into a few training data with some particular patterns and flip the labels but has a high attack success rate. However, this attack will leave a trace of backdoor in the activation filters layers. [28] proposed clean-label attack, the two approaches: GAN-based and adversarial-attack-based are introduced in paper [28]. Compared with the BadNets attack, the adversary only needs to poison a small percentage of a single class, but the attack ability is strong. Interestingly, [29] proposed an optimization-based method to generated the poisoned data, which only needs to poison the target class with one image and can reach well-performance on ASR. Although these early methods have high ASR, they can be detected with some standard detection methods([30]–[32]). However, [33] designed an embedding algorithm, which can bypass the detection approaches. They added a discriminator to distinguish the poisoned data and clean data, then update the parameters of the model with the loss of discriminator, aiming to confuse the discriminator. The detection approaches will fail because the feature representation of the poisoned and clean data is mixed when the discriminator cannot distinguish them. [34] presented the backdoor attack with dynamic triggers, the similar idea is claimed in [35]. They used a generator network to generate the triggers according to the inputs. The triggers of different inputs are also different. However, the drawbacks of the approaches in [34], [35] are that the triggers are still apparent for humans and can be detected by then visual detection. [22] proposed a novel attack approach *Refool* that used physical reflection properties to implant backdoors. The adversary will choose some images from candidate images subsets, and these images are inserted into the target images as the triggers through the reflection algorithms in [22]. [23] proposed attack approach *WaNet* that using a small and smooth warping field to generate poisoned examples. The poisoned examples of WaNet are not inserted into the visual patterns but with the small unnoticeable distortion. Besides, [18] firstly released semantic attack to backdoor federated learning. [36], [37] also introduced the similar ideas of semantic backdoor attacks. The semantic backdoor attack is stealthy for humans because the triggers are more realistic and easily ignored.

Moreover, there exist some attacks to modify the models. [38] first explored the attack based on modifying the model parameters directly without training with poisoned data. [39] injected the backdoor through inverting bits on the quantization network. [40] also proposed a similar attack method. At present, [41] showed how to backdoor attack contrast learning. They listed many enlightening findings in their paper and claimed that poisoning only 0.005% can cause the model to misclassify. In the future, backdoor attacks with self-supervised learning will be a promising research field.

B. DEFENSE

Although the backdoor attack has a certain stealthiness and is challenging to be detected by humans, most of the backdoor attacks will still leave a trace to be detected. [30] proposed that the spectrum of the feature represents covariance can be used to detect the backdoor. [31] found that the feature representations of the poisoned data and clean data are different in that they could be clustered into two clusters, one is poisoned, and the other is clean. Based on an intuition that the triggers are usually small so that [32] used the optimization method to get reverse engineer trigger for each target label and measured the $L1$ norm of the triggers. They defined anomaly index, the absolute deviation of the data point divided by the median of their absolute deviations, as the detection

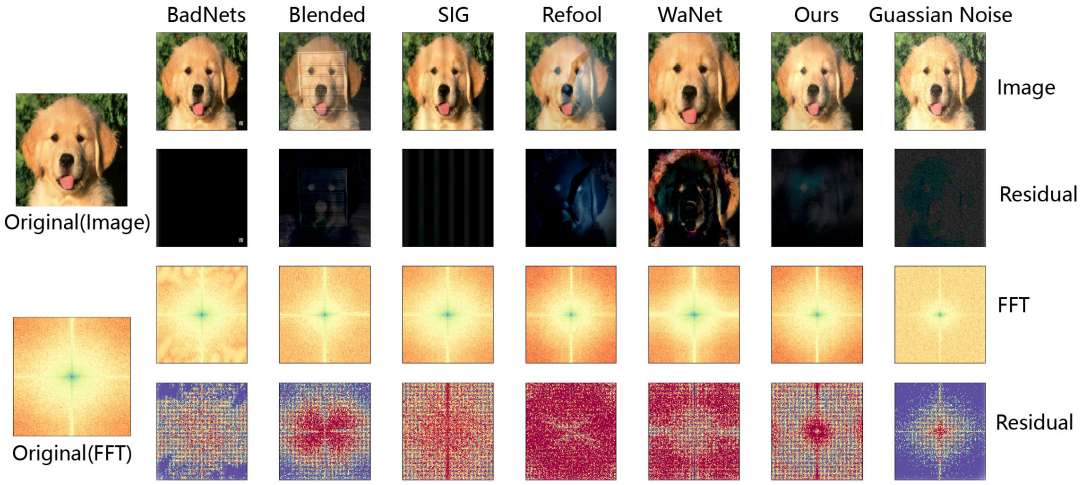

FIGURE 1. Comparison among different backdoor attacks. The leftmost is the original images and the frequency domain image with Fast Fourier Transform(FFT). The top row shows the poisoned examples generated by BadNets([9]), Blended([20]), SIG([21]), Refool([22]), WaNet([23]) and our method. The last one of that is the image adding natural noise, taking it as a natural condition. The second row shows the residual map, following [23]. The poisoned examples generated by the first four methods are unnatural. Both the WaNet method and our method are not easy to detect by humans, and our method has fewer changes to the original image. We transform images through FFT to show more details because the high-frequency domain can show the changes in detail. It can be seen that the images change a little under our method in the high frequency, which highlights the stealthy of our approach.

criterion. [42] proposed a method utilizing a likelihood-ratio test to analyze the feature representations and identify the legitimate image. [43] perturbed the inputs by superimposing various image patterns and observing the entropies of inputs. They claimed that the entropies of poisoned examples are lower than that of the clean. Recently, [44] introduced to compute adversarial perturbation for each input and utilize the difference between the perturbations for the clean examples and the poisoned examples to detect. Besides, [45] studied the dissimilarities between poisoned examples and clean examples under the frequency domain and proposed a feasible detection method. In addition to these detection methods, there also exist many approaches that mitigating the attack ability. [46] proposed to use neuron pruning to mitigate backdoor. [32] proposed Neural Cleanse with three approaches: input filters, neuron pruning, and unlearning. These three methods provided some enlightening ideas for the defense against backdoor attacks. [47] pointed out the vulnerability of triggers and demonstrated that transformation methods could defend against backdoor attacks. [48] showed that adversarial training could be used to enhance the robustness of the model to mitigate attacks. Besides, [49] proved that using $L2$ regularization could mitigate backdoor attacks and the tight parameters are the key to this defense approach. [50] studied a lot of specific attack methods and defense methods in detail, and listed the advantages and disadvantages of these methods, puts forward some incisive conclusions. Therefore, it is suggested that readers can read through this paper. Following the [50], we also suggest the researchers can have a better understanding about some resources on https://pages.nist.gov/trojai/.

III. PRELIMINARY

A. BACKDOOR ATTACK

In the backdoor attack, the adversary can get to the clean training dataset D and modify the clean dataset by injecting some poisoned data P following $D' = D \cup P$, the backdoor model f' can be obtained through training with poisoned dataset D'. The backdoor model f' has regular classification performance on the clean dataset but classifies the poisoned data with triggers into the labels that the adversary wants. This attack mode is often a targeted attack. The poisoned example x' is normally written as $x' = x \oplus \mathcal{T}$, where $\mathcal{T}$ is the backdoor trigger. In this paper, the generation of poisoned data is following as:

$$x' = G(x) = clip((1-\alpha)\cdot x + \alpha \cdot \mathcal{T}, x_{min}, x_{max}) \quad (1)$$

where $G(\cdot)$ is the generation function, $\alpha \in [0, 1]$ is a trade-off hyper-parameter, x' is the poisoned example. And x_{min}, x_{max} are the minimum and maximum values of images x. Attack success rate(ASR) is a crucial index to measure the effect of the backdoor attack, which means the success rate of classifying the poisoned data into the target class. It is as follows:

$$ASR = \frac{\sum_{x,y \in D_{test}} [f'(G(x)) = y' | y \neq y']}{N_{D_{test}}} \quad (2)$$

where y' is target class, and $N_{D_{test}}$ is the number of dataset D_{test}.

B. PROCEDURAL NOISE

In Computer Graph field, Noise is defined as *the random number generator of computer graphics*([51]). The function

of noise is always characterized by the power spectrum. If we can manipulate the power spectrum of the noise, we can model the noise in a highly structured texture. Procedural noise is a special class of noise that has rich texture features. Meanwhile, it has a rich mathematical theory to follow and easily interpret. There are many methods to generate procedural noise, which are classified into the following:

- **Lattice**
  - **Perlin** One of the famous procedural noise generation algorithms is the Perlin noise algorithm([52]). To generated Perlin noise, firstly, we define a lattice structure and compute the pseudo-random gradient vector of the vertices of each lattice. Secondly, we compute the distance between point (x, y) and its adjacent lattice vertices, then dot them with the gradient of vertices. Finally, we compute the weight sums by using the relaxation curve. The relaxation curve function is often as following function: $y = 3t^2 - 2t^3$ or $y = 6^5 - 15t^4 + 10t^3$.
  - **Simplex** Simplex noise is an improvement on Perlin noise([53]), the complexity of Perlin noise is $O(2^n)$, but Simplex noise is $O(n^2)$ and the lattice type of Simplex noise is also different from that of Perlin noise.
- **Convolution**
  - **Gabor** Gabor noise is a sparse convolution noise with the Gabor kernel g([54]) as following:

$$g(x, y) = Ke^{-\pi a^2(x^2+y^2)} \times \cos\left[2\pi\lambda\left(x\cos\omega_0 + y\sin\omega_0\right)\right] \quad (3)$$

    where K and a are the magnitude and inverse width of the Gaussian kernel, λ, ω_0 are the period and orientation. For a point (x, y), the Gabor noise $N_{x,y}$ in (x, y) is computed as follows:

$$N(x, y) = \sum w_i g\left(K_i, a_i, \lambda_i, \omega_{0,i}; x - x_i, y - y_i\right) \quad (4)$$

    where w_i is the random weights, (x_i, y_i) are the random positions, K_i, a_i, λ_i and $w_{0,i}$ are the parameters of the different Gabor kernels.
- **Point**
  - **Worley** [55] proposed a function-based Voronoi graph to generate procedural noise. This noise has a lattice texture, and it can produce textured surfaces resembling flagstone-like tiled areas, organic crusty skin([55]).
  - **Voronoi** Voronoi noise is based on Worley noise, whose edges are much smoother and the textures of which are more natural.

IV. PROCEDURAL NOISE ATTACK

A. ATTACK MOTIVATION

How to make the design of the trigger more invisible? Comparing with local triggers([9], [28]), global triggers([22], [23]) more stealthy and unnoticeable. The triggers of [22], [23] are global and show the powerful attack capability and stealthiness(Fig.1). Therefore, there exists an intuition that designing a global trigger, which is easy to be learned for deep learning. It can be interpreted as following:

The texture is an essential feature of the image. If we modify the texture feature of inputs slightly and teach the model to learn this new texture feature, can we make the model remember this textured pattern and take it as a trigger?

Procedural noise has strong texture features, and it is often used in terrain generation, computer graphics, and other fields([51]). According to intuition, we can add an image with different types of procedural noise by Equation.1. To show the effectiveness in the worst case, we choose a picture whose background is pure color, Fig.2 shows the demo of the image with procedural noise. The fusion scalar α in Equation.1 is set as 0.2. We can find that images with the Perlin noise, Simplex noise, and Voronoi noise are still natural even in the worst case. Although there exist some shadow regions in the image, they are often ignored by the vision system of humans.

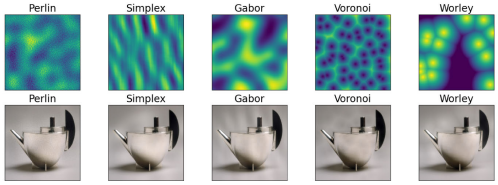

FIGURE 2. The images with different procedural noise. The top row is the procedural noise, and the bottom row is the images adding the corresponding procedural noise. We set the fusion scalar as 0.2.

[56] proposed to use procedural noise as the universal adversarial perturbation for the black-box adversarial attack. They claimed that the procedural noise has high ASR to cheat the classifiers because the procedural noise and the other existing Universal Adversarial Perturbations (UAPs) are similar. [57] claimed that the adversarial perturbations could change the features of inputs. Therefore, we can conclude that the procedural noise, similar to UAP, can change the features of inputs and attack the model. Our motivation is to utilize the property that changing the feature to make small perturbations to the target object. In the backdoor attack, we use procedural noise to change the features of inputs and make the model learned that changed features as the potential triggers. In the adversarial attack, procedural noise changes the features of inputs and confuses the model to get an incorrect result. There is a significant gap between the backdoor attack and the adversarial attack with procedural noise. Next, we will show our attack approach in detail.

B. ATTACK APPROACH

Our attack method can be divided into the following steps:

1) Generate procedural noise. In this paper, we introduce three different categories of procedural noise. As we can see from Fig.2, we can find that the different types

of procedural noise in the same category are similar. Therefore, we can choose one from each category as representative. In this paper, Perlin noise, Gabor noise, and Worley are chosen to evaluate the attack effectiveness. Meanwhile, we will adjust the procedural noise and tune it for practical purposes to make our noise challenging to detect.

2) Train a clean model to gain the attention images of GradCAM([58]). If we add noise directly to the large-scale images, the backdoor model will remember the texture features of the whole procedural noise images. In this case, the attention of the GradCAM is abnormal, and it is easy to detect by using the approaches of XAI. Therefore, we choose the attention image of the model as a mask for the procedural noise images, and the noise to be injected into the image is the procedural noise dot product with the mask. It will be explained in the next itemize in detail.
3) Fuse image and noise. We define the procedural noise as $\mathcal{I}_p$, the attention image as $\mathcal{I}_a$. To mitigate the effect of the Mach Bach, we use the sin function to fuse the noise and image. The mask is defined as follows:

$$mask = dot(sin(\frac{\pi}{2}\mathcal{I}_a), \mathcal{I}_p) \tag{5}$$

where the *dot* is the dot product, next, we will use the mask noise as the backdoor trigger $\mathcal{T}$.
4) Train the backdoor model. The whole training process can be seen in Fig.3.

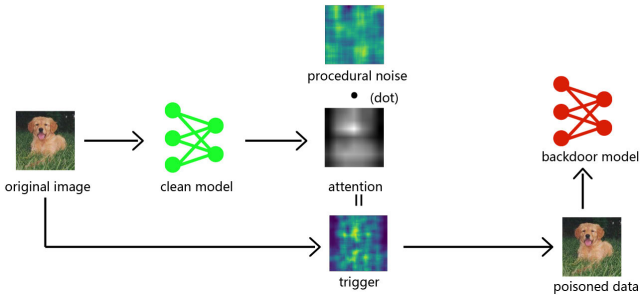

FIGURE 3. An overview of our attack approach. The whole training process for the backdoor model follows: Firstly, we extract the attention image for the original image through GradCAM methods. Secondly, we randomly generate the procedural noise and dot product with the attention image to produce the trigger. Finally, we generate the poisoned example following Equation.1 and train model with such poisoned example.

V. EXPERIMENT

A. EXPERIMENT SETUP

In this paper, we evaluate our attack approach on four different datasets: CIFAR-10, GTSRB, CelebA, and ImageNet12. These four datasets are the benchmark datasets in the Computer Vision(CV) field. The evaluation results can provide fair and trustworthy comparisons. For the CelebA dataset, there exist 40 attributions, and we choose three attributions of them and split the CelebA dataset into eight classes to do multi-class classification. For ImageNet12, we choose 12 classes from the total ImageNet dataset following the experiments settings in [22].

- **CIFAR-10** CIFAR-10 dataset has 10 classes, whose images are 3 channels. The size of images from the CIFAR-10 dataset is $32\times32\times3$. The training set consists of 50K examples, and the testing set consists of 10K examples.
- **GTSRB** GTSRB dataset is a multi-class, single-image classification dataset, which has 43 classes, and more than 50,000 images in total. The size of images from GTSRB is from small to large, and we resize them into the scale $32 \times 32 \times 3$.
- **CelebA** CelebA is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. In this paper, we resize them into the scale $64 \times 64 \times 3$.
- **ImageNet12** ImageNet12 is a subset of the ImageNet dataset. The entire training set consists of 12492 images, and the testing set consists of 3132 images. In this paper, we resize them into the scale $224 \times 224 \times 3$.

In this paper, we use ResNet18 to train CIFAR-10, GTSRB, and CelebA datasets and use DenseNet121 to train the ImageNet12 dataset. The optimizer is Adam optimizer with an initial learning rate of 0.001. Besides, the fusion scalar α is 0.3 for CIFAR-10 and GTSRB, and that for CelebA, ImageNet12 is 0.2. The poison ratio is 0.1.

B. ATTACK EXPERIMENTS

This paper lists three types of procedural noise, and each type of procedural noise has multi-hyperparameters for its functions. Naturally, there will raise some questions:

- *Do the hyperparameters of procedural noise have effects on the attack efficiency?*
- *Do different types of procedural noise have effects on the attack efficiency?*

There exists more than one hyperparameter for procedural noise. To simplify the question, we choose one hyperparameter as the controlled variable, and we set different hyperparameters values for Perlin noise(lattice-based method), Garbor noise(convolution-based method), and Worley noise(point-based method). For Perlin noise, the period is the critical factor to control the noise. If the period is small, the scale of Perlin noise is much smaller (Fig.4 top). Therefore, we choose the period p as the controlled variable. For Gabor noise, we choose λ of Equation.3 as the controlled variable. The Gabor noise with different λ and the images with these noises are shown in the middle of Fig.4. For the Worley noise, the number of lattice n is the key parameter. We generated the Worley noise with different n, and the images with the Worley noise are shown at the bottom of Fig.4.

The period p for Perlin noise is set from 10 to 100, the period λ for Gabor noise is set from 5 to 50, and the number of points n for Worley is set from 2 to 20. The other hyperparameters are stable that following the [56]. The fusion

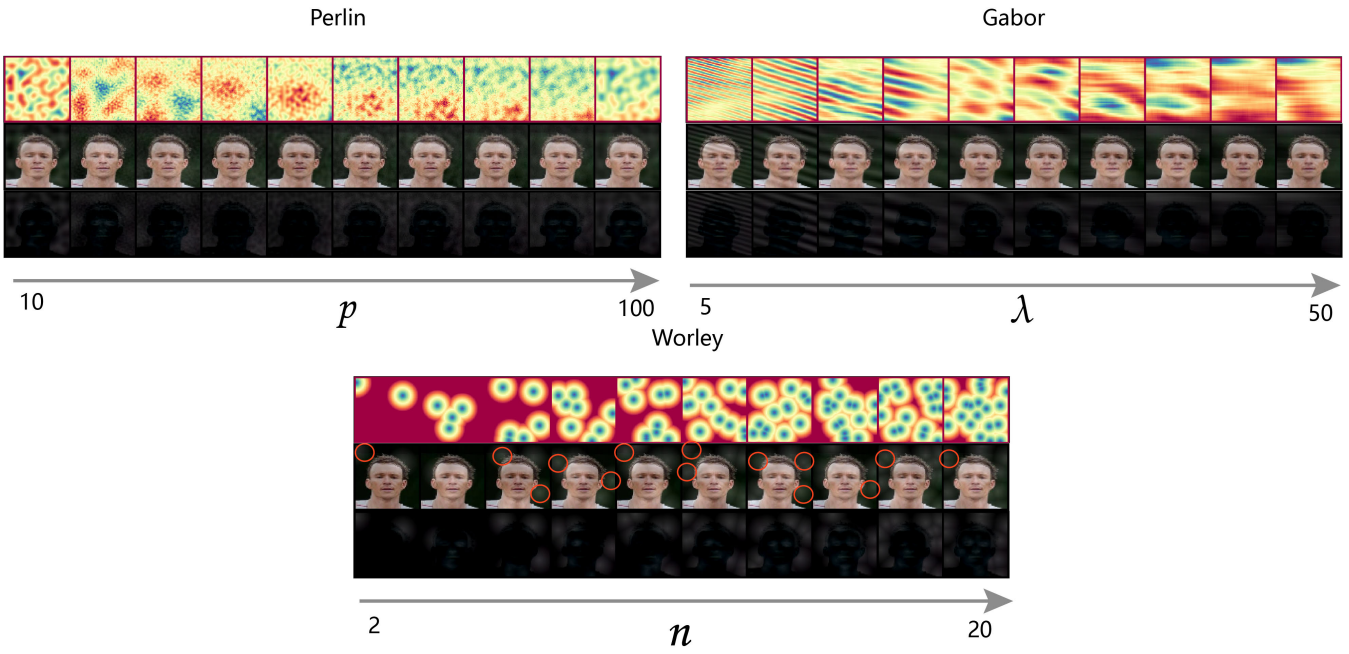

FIGURE 4. The demo of the images with different procedural noise. We choose Perlin noise, Gabor noise, and Worley noise. The first row of each sub-figure is the corresponding procedural noise, the second row is the images with procedural noise, and the third row is the residual image. We set fusion scalar as 0.2, and the g in Equation.5 in the $1_{h\times w}$, where h, w are the height and width of image. For the large-scale dataset(ImageNet12), the g is usually set as the pixel values of the attention image. Especially for images with the Worley noise, we mark some regions with red circles that may be used as a basis for judging the backdoor attack.

scaler is set as 0.2. The demo of the images with different procedural noise is shown in Fig.4. As we can see from Fig.4, the perturbations of Perlin noise are more subtle and more fine-grained. Although the periods are from small to large, the differences among these periods are not noticeable. The distortions are substantial enough for the Gabor noise when λ is small, and the distortions will decrease when the λ becomes large. When the λ is 50, the bandwidth of the procedural noise image is large so that the changes in the neighbor region are not drastic, and we can find the image with the Gabor noise under the λ is 50 are natural. For the Worley noise, it is clear that the Worley noise is like clusters, making corresponding target image regions turn shallow and become abnormal. To highlight this case, we mark the shadow regions with red circles that may leave the trace for humans to be detected.

To evaluate the attack ability of the procedural noise under different hyperparameters, we compute the accuracy and ASR. The evaluation results are shown in Table.1. As we can see, each accuracy of the backdoor model is similar to the standard clean model, and the ASR is also high enough to prove the powerful attack ability. Meanwhile, we can find few differences for the backdoor with the same procedural noise but different hyperparameters. The different types of procedural noise have similar performances. Therefore, we can conclude that *the hyperparameters of procedural noise have few effects on the attack efficiency, and they all have high accuracy and ASR.* Based on the conclusion, we can have a more stealthy and fine-grained backdoor attack, which is not easy to detect.

C. DEFENSE EXPERIMENTS

To evaluate if our attack approach can attack against the most defense methods, we choose Activate Clustering([31]), Spectral Signatures([30]), Fine-Pruning([46]), Neural Cleanse([32]), STRIPS([43]) and GradCAM Visualization as baseline defense approaches. The main introductions of these defense approaches are following:

- **Activate Clustering** Activate clustering is a detection method to defense against backdoor attacks. [31] found the feature representations can be clustered into two clusters. One is a poisoned cluster. The other one is a clean cluster. The last hidden layer can reflect the high-level features of the inputs([50]). If there exist differences among the high-level features, the feature representations can interpret these differences. Therefore, activate clustering utilizes this property to detect the poisoned examples.
- **Spectral Signatures** The spectral signatures method is also a detection method to detect whether the inputs belong to poisoned data. [30] claimed that the backdoor model would leave the trace for the poisoned data. They used singular value decomposition to decompose the latent representations and computed the outlier score. If the outlier score is high, the input will be judged to be a poisoned example. Otherwise, the input is a clean example.
- **Fine-Pruning** Pruning technology is a valuable technology in deep learning. Fine-pruning is to prune the activated neurons to mitigated the backdoor attack carefully.

TABLE 1. Accuracy and ASR for the procedural noise with different hyperparameters.

	Perlin(p)																			
	10		20		30		40		50		60		70		80		90		100	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR-10	92.67	99.12	91.89	99.23	92.42	99.17	92.79	98.99	91.58	99.23	91.87	99.42	92.28	99.31	91.92	99.25	91.87	99.17	91.99	98.99
GTSRB	96.18	99.12	95.99	98.21	96.16	98.95	95.93	99.10	95.08	99.23	95.63	99.10	95.99	98.78	96.23	98.99	96.12	99.13	96.02	99.43
CelebA	78.53	99.86	79.24	99.54	79.35	99.82	79.49	99.91	80.01	99.87	79.48	99.89	78.99	99.92	79.72	99.91	79.91	99.85	79.25	99.79
ImageNet-12	96.57	99.23	96.61	99.12	97.01	99.15	96.78	98.89	96.73	99.38	96.32	99.10	95.93	99.12	96.73	99.01	97.01	98.75	96.88	99.12
	Gabor(λ)																			
	5		10		15		20		25		30		35		40		45		50	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR-10	92.11	99.49	91.92	99.32	91.69	99.12	91.44	99.23	91.14	99.33	91.16	98.76	91.32	99.22	90.64	98.19	91.26	98.98	91.29	99.19
GTSRB	95.85	99.23	95.34	99.14	95.78	99.12	94.98	98.88	95.09	98.99	95.15	99.22	95.55	99.33	94.92	99.27	95.32	99.11	94.99	99.27
CelebA	78.99	99.77	79.12	99.32	79.46	99.64	78.34	99.57	79.56	99.78	80.02	99.83	79.56	99.24	80.03	99.76	79.99	99.32	79.64	99.41
ImageNet-12	96.78	99.58	96.83	99.24	97.05	99.16	96.89	99.33	97.23	99.54	97.26	98.72	96.18	99.53	97.12	99.38	97.66	98.97	97.78	99.41
	Worley(n)																			
	2		4		6		8		10		12		14		16		18		20	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR-10	90.76	99.11	91.39	99.26	91.52	98.23	91.32	98.01	89.48	98.86	89.91	96.25	89.88	98.34	90.21	98.71	90.44	98.99	90.11	99.00
GTSRB	95.23	99.82	95.27	99.12	95.39	99.22	95.21	99.27	96.08	98.98	95.81	99.41	96.15	99.21	95.34	99.70	95.88	98.34	95.13	99.84
CelebA	78.25	99.35	78.44	99.26	79.01	98.73	78.82	98.93	78.48	99.75	79.22	99.48	79.34	98.82	78.46	99.17	78.99	99.24	79.39	99.32
ImageNet-12	96.57	99.61	97.01	99.23	96.17	99.65	97.06	98.93	96.34	99.04	96.59	99.35	96.99	99.21	96.25	98.77	97.25	99.33	96.55	98.35

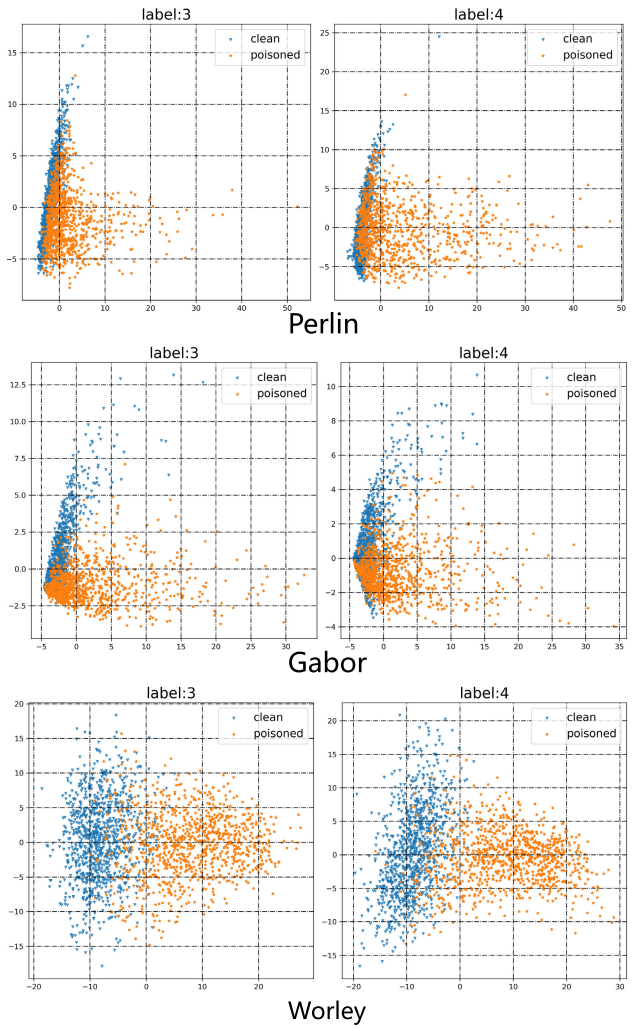

FIGURE 5. The representations of the poisoned examples and clean examples. We use PCA to represent the last hidden layer and K-means to cluster the poisoned clusters and clean clusters.

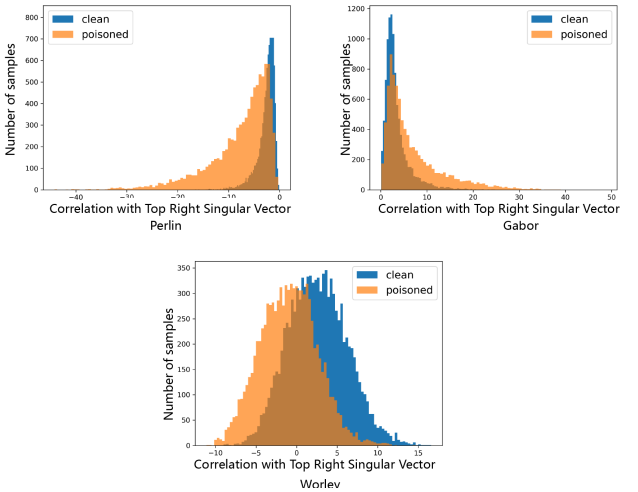

FIGURE 6. The detection for the three types of procedural noise attack on the CIFAR-10 dataset through spectral signatures.

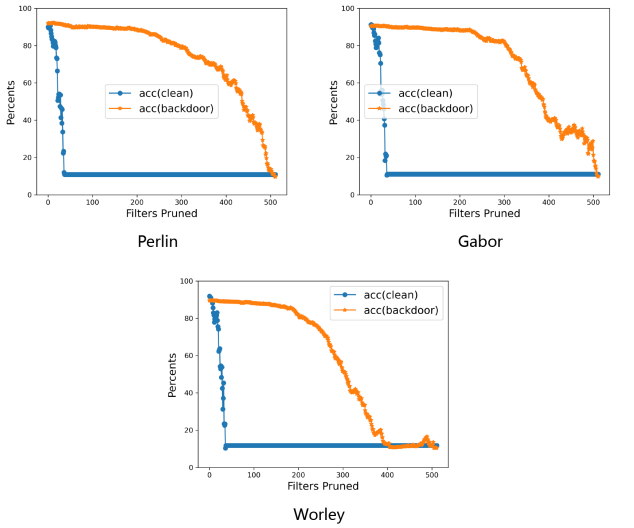

FIGURE 7. Fine-Prunning on CIFAR-10. The x-axis is the number of pruned filters, and the y-axis is accuracy.

Inspiring the neurons will be activated when meeting the poisoned trigger([9]), [46] proposed to prune the backdoor neurons then restore the model performance. However, this method will degrade the accuracy of the model, and pruning each neuron will have high compute complexity.

- **Neural Cleanse** Neural Cleanse is also a detection approach against backdoor attacks. [32] used reverse

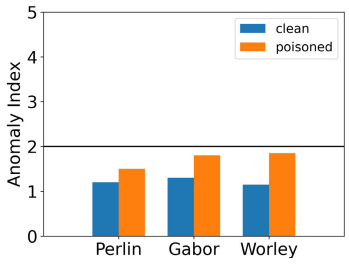

FIGURE 8. The anomaly index of the clean data and poisoned data on the dataset CIFAR-10.

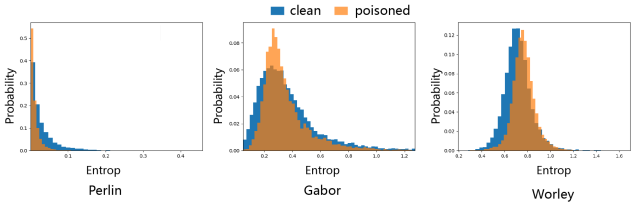

FIGURE 9. The entropies of poisoned examples and clean examples on CIFAR-10.

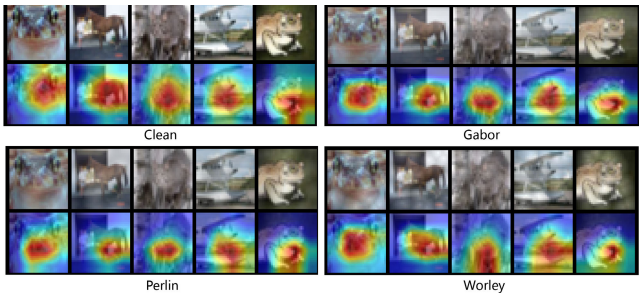

FIGURE 10. The activation heatmaps of the clean model and backdoor models on the dataset CIFAR-10.

engineer to get the potential trigger for each label and computed the $L1$ norm for each trigger. This input will be a poisoned example if the potential trigger deviates from the median absolute deviations of all potential triggers. [32] defined anomaly index as the detection criterion, and the anomaly index is always set as 2.

- **STRIP** STRIP is also a baseline detection method. The intuition is that the prediction of clean input under perturbations will have high uncertainty. However, the uncertainty of predicting poisoned examples will be low, which has a strong hijacking effect to ensure the attack success rate. [43] claimed that this method is input-agnostic and can detect the poisoned examples with the threshold of entropy.
- **GradCAM Visualization**The deep learning model often lacks the explainability of decisions. However, Explained AI(XAI) provides practical tools to comprehend the deep learning model. The supplement decisions of XAI make it possible to detect backdoor attacks. For a poisoned input, the defender will find the anomalies if the attention of the model is not suitable. Therefore, we will utilize GradCAM to visualize the attention regions of the model and illustrate that the attention of the backdoor model is similar to that of the clean model.

FIGURE 11. The original image with the generated images by 17 different corruption methods.

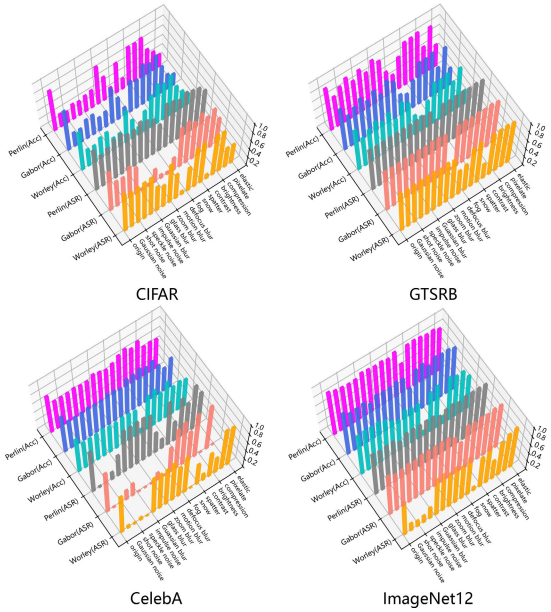

FIGURE 12. The accuracy and ASR of backdoor models with three types of procedural noise.

TABLE 2. Activate clustering detection results on CIFAR-10.

		0	1	2	3	4	5	6	8	9
Perlin	Acc	52.59	54.53	53.12	55.14	56.54	56.44	56.52	54.99	55.93
	F1 Score	55.80	68.14	71.69	70.35	72.08	72.00	71.88	69.44	71.10
		0	1	2	3	4	5	6	8	9
Gabor	Acc	61.37	62.79	64.24	64.72	65.13	63.24	65.48	63.19	62.53
	F1 Score	71.73	73.83	75.82	76.93	74.22	74.45	77.22	76.29	74.06
		0	1	2	3	4	5	6	8	9
Worley	Acc	99.53	89.39	99.10	95.31	99.62	98.89	94.23	99.14	93.12
	F1 Score	92.33	88.74	86.66	95.35	88.86	88.56	95.30	95.65	94.58

To have the best stealthiness, we set the period p of Perlin noise as 100, the λ of Gabor noise as 50, the number of points n of Worley noise as 2. In the experiment section, we mainly discuss the results of the CIFAR-10, and the other results of other datasets are shown in Appendix. Moreover, the conclusions about CIFAR-10 also fit the other datasets.

1) ACTIVATE CLUSTERING

Following the experiment settings from [31], we plot the distribution of the poisoned examples(three types of procedural noise) and clean examples on the CIFAR-10 dataset with PCA(Fig.5). As we can see from Fig.5, the clean examples(blue points) and the poisoned examples(orange

FIGURE 13. The poisoned data on the other datasets.

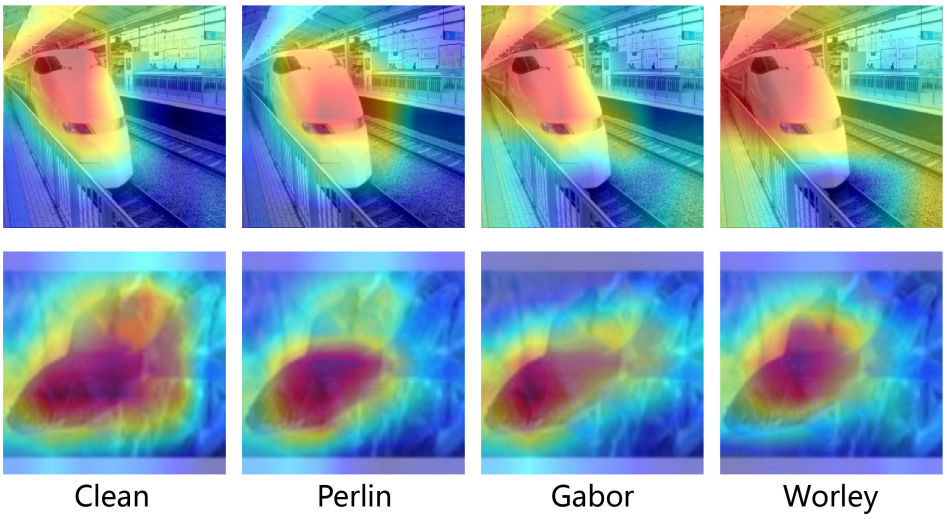

FIGURE 14. GradCam.

points) are mixed under the backdoor attack with Perlin noise and Gabor noise, except for the Worley noise. The results mean that the poisoned data with Worley noise can be detected by the activate clustering approach, but the poisoned data with Perlin noise and Gabor noise can bypass the detection. To evaluate the detection effectiveness, Table.2 shows the accuracy and F1-score of the detection on the clean data and poisoned data under different labels. The Table.2 also prove the conclusion that we get from Fig. 5.

2) SPECTRAL SIGNATURES

For the spectral signatures([30]), we compute the top right singular vectors of the feature representations. The result is shown in Fig.6. The x-axis shows the corresponding eigenvalue, and the y-axis shows the number of examples with the corresponding eigenvalue. In contrast to the demonstrations in [30], the eigenvalues of the poisoned examples with three types of procedural noise are similar to the clean examples. Even for the Worley noise, it can also bypass the detection. For the detector, it is hard to separate the poisoned examples from the clean examples. Therefore, the defender cannot mitigate the backdoor attack by throwing away the poisoned data.

3) FINE-PRUNING

Fine-Pruning mainly focuses on neuron analyses. For a given specific layer, the last hidden layer is chosen as the specific layer. We test Fine-Pruning on our models with different procedural noise and plot the accuracy of clean data with the backdoor model and clean model. As we can see from Fig.7, there exist no points that the accuracy of the clean is higher than the backdoor, meaning that our approach can bypass Fine-pruning, and it is difficult for Fine-Pruning to mitigate the attack.

4) NEURAL CLEANSE

[50] claimed that the Neural Cleanse is less effective with the increased trigger size. The results of our experiment on CIFAR-10 prove the demonstration from [50]. As we can

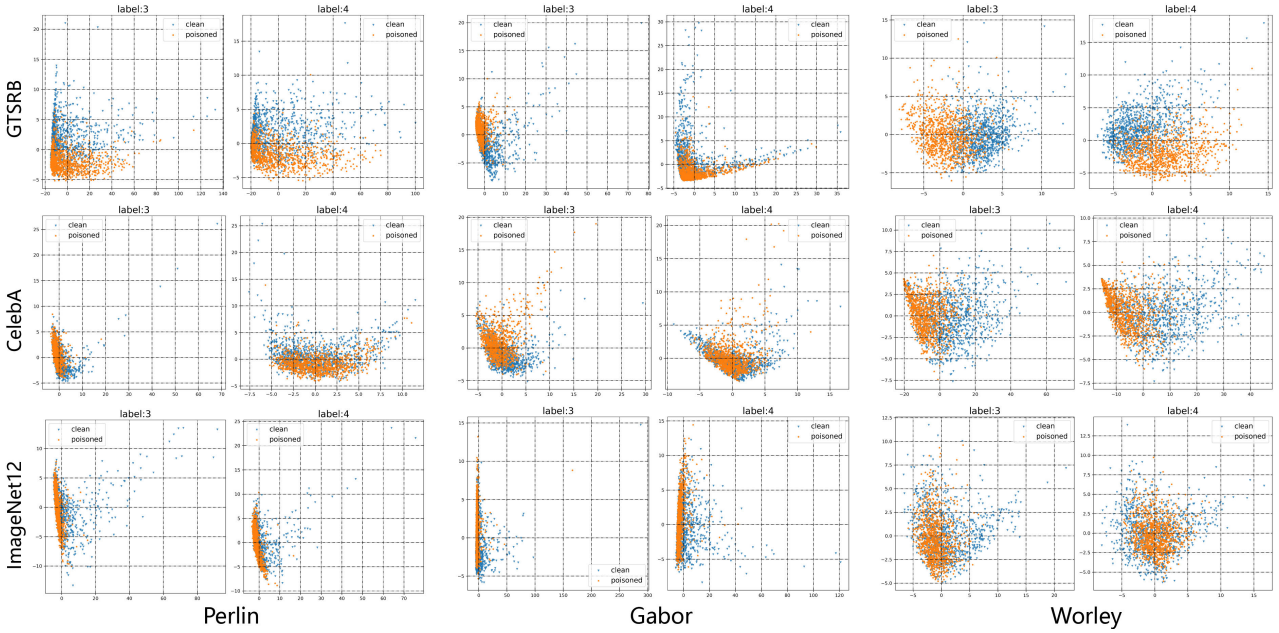

FIGURE 15. Activate Clustering. The feature representations of the clean data and poisoned data.

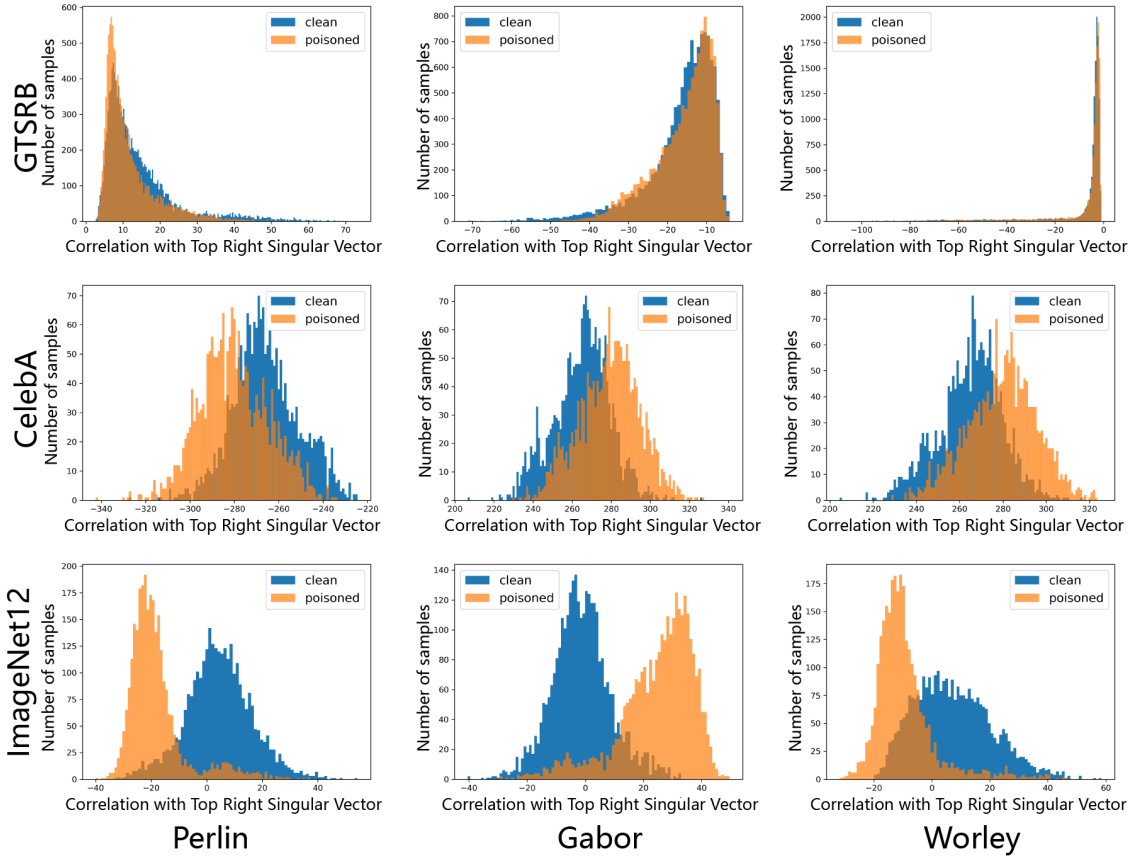

FIGURE 16. Spectral signatures.

see from Fig.8, the anomaly index of the poisoned data with procedural noise is all under threshold 2(The recommended value in [32]), which means that our approach can bypass the Neural Cleanse.

5) STRIP

In the evaluation with STRIP, we randomly choose 100 clean images from different classes to superimpose them into the test images and compute the entropies of inputs. Fig.9

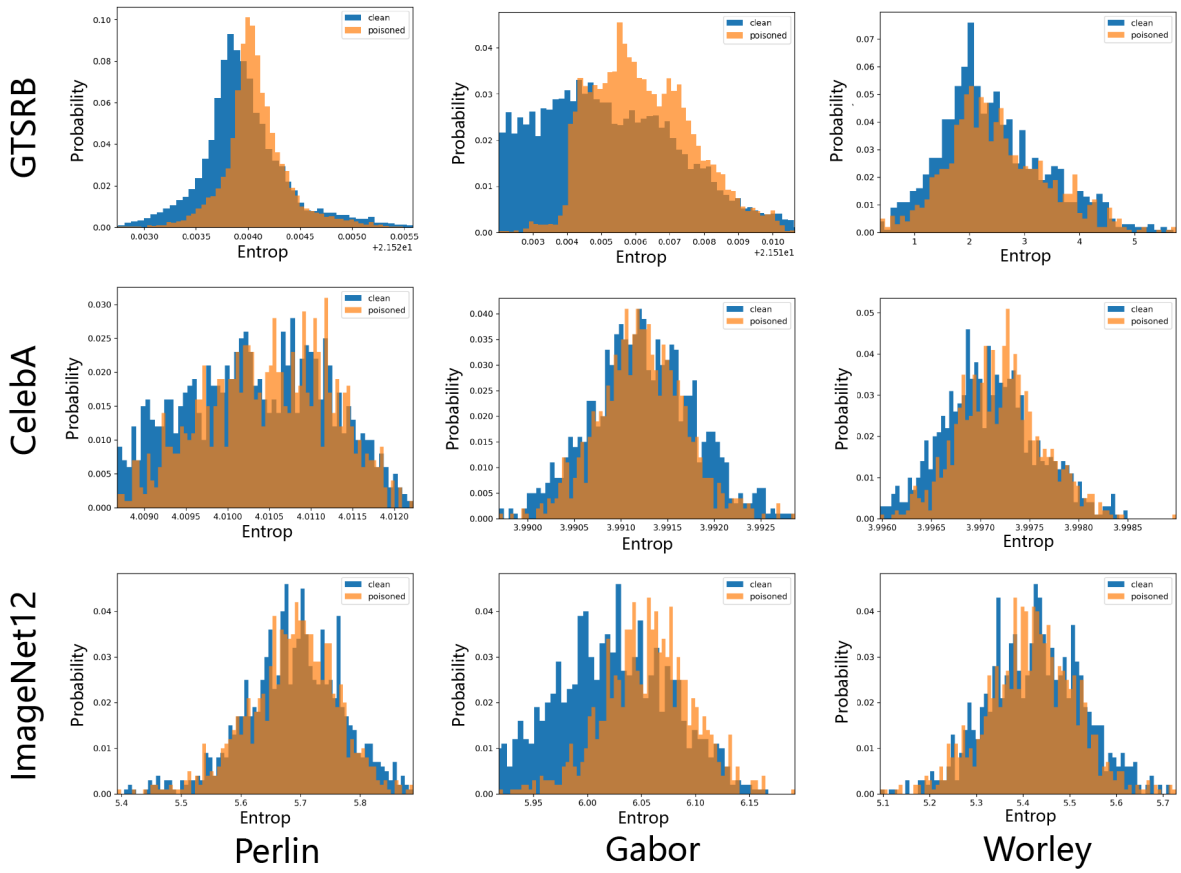

FIGURE 17. STRIP.

shows the entropies of the poisoned examples and clean examples. We can see that the entropies of clean examples and poisoned examples have similar distributions. According to the principle of STRIP, the poisoned examples usually have low entropies, and the clean examples usually have high entropies, which fails for poisoned data with procedural noise. The backdoor model generated by our approach behaves like a benign model.

6) GradCAM VISUALIZATION

We randomly choose five images from the CIFAR-10 dataset and visualize the activation heatmaps of the clean model with backdoor models in Fig.10. As can be seen, the heatmaps of the clean model and the backdoor models are similar and there exist few differences among them. Although the poisoned data is injected into the trigger, procedural noise, the heatmaps do not change a lot, which means that our approach can bypass the GradCAM visualization detection.

D. ATTACK ROBUSTNESS

In practice, the backdoor triggers will become fragile for disturbances, which will cause ASR to decrease through some simple transformations or corruption methods. However, humans are not confused about these transformations or corruption methods. There will exist some limitations for the adversary to deploy the backdoor attacks in the physical scene. We evaluate our attack approach with different corruption methods to test the robustness. We choose 17 corruption methods following [59]. These seventeen corruption methods include Gaussian noise, shot noise, speckle noise, impulse noise, Gaussian blur, glass blur, zoom blur, motion blur, defocus blur, fog, snow, spatter, contrast, brightness, jpeg compression, pixelate, and elastic transform. We choose an image from GTSRB and generate different images by 17 different corruption methods shown in Fig.11.

We plot the accuracy and ASR of backdoor models with three types of procedural noise under the corruption methods in Fig.12. On the CIFAR-10 dataset, we can find that the accuracy of the backdoor models decreases under the most corruptions, and these models have a similar performance of accuracy. However, there exist gaps among the ASR of the backdoor models with different procedural. The Perlin noise outperforms other procedural noise, and the different types of procedural noise have different sensitivities for the corruption methods. The ASR of Gabor decreases the most with noise the Gaussian blur, glass blur, motion blur, defocus blur, fog, snow. The Worley noise is affected by the defocus

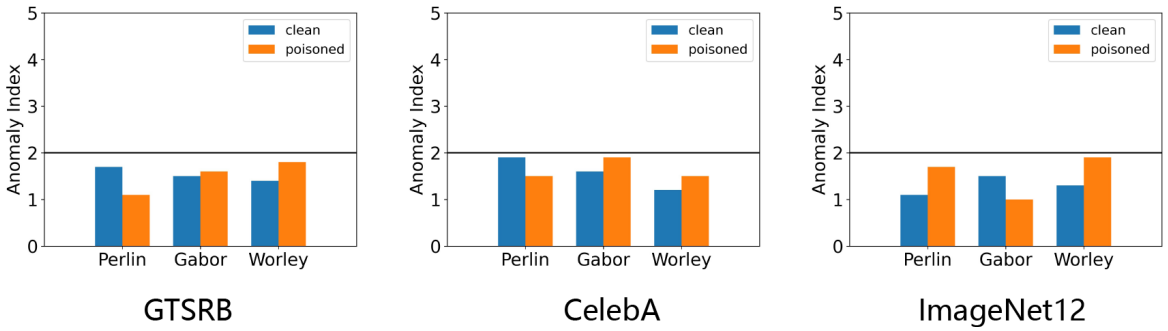

FIGURE 18. Neural cleanse.

blur and contrast. In other datasets, the conclusion about the sensitivities of the backdoor models is also established. In general, the Perlin noise is much more robust than other procedural noise. The backdoor models with Gabor noise and Worley noise are easily affected by corruption methods, which are not entirely the same. Besides, we can find that our approach is robust for most corruption methods, which means our approach can be applied to the physical scenario with the adaptive ability. In fact, the backdoor attack can affect the deep learning models' downstream tasks and can also attack the applications about deep learning, such as face recognition and autonomous driving. Therefore, the defense approach against such attacks is necessary, and it will be the future works for us.

VI. CONCLUSION

In this paper, we introduce a novel backdoor attack approach that perturbs the target images with procedural noise, whose triggers are powerful and stealthy. We study three different types of procedural noise and prove that our attack approach can bypass most defense methods. Besides, the experiment with corruption also claims that our approach has strong robustness for many corruption methods, and they can be applied in practice. In the future, we will develop the defense framework to defend some related backdoor attacks such as Refool([22]), WaNet([23]) and our attack approach. Moreover, we will also investigate the theory of the backdoor attack to have a deeper understanding.

APPENDIX
THE EXPERIMENTS ON THE OTHER DATASETS

See Figs. 13–18.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[2] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, pp. 681–694, Nov. 2020.

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[4] I. D. Apostolopoulos and T. Bessiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, 2020.

[5] C. Zheng, X. bo Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang, "Deep learning-based detection for COVID-19 from chest ct using weak label," *MedRxiv*, Jan. 2020.

[6] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110059.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*. [Online]. Available: http://arxiv.org/abs/1312.6199

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, pp. 1–11, Dec. 2015.

[9] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," 2017, *arXiv:1708.06733*. [Online]. Available: http://arxiv.org/abs/1708.06733

[10] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. NDSS*, 2018, pp. 1–17.

[11] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1332–1333.

[12] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," 2020, *arXiv:2007.14321*. [Online]. Available: http://arxiv.org/abs/2007.14321

[13] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1528–1540.

[14] G. Qi, L. Gong, Y. Song, K. Ma, and Y. Zheng, "Stabilized medical image attacks," 2021, *arXiv:2103.05232*. [Online]. Available: http://arxiv.org/abs/2103.05232

[15] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: Review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, Jun. 2019, Art. no. e01802.

[16] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 997–1005.

[17] H. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.

[18] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. AISTATS*, 2020, pp. 2938–2948.

[19] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," 2020, *arXiv:2007.05084*. [Online]. Available: http://arxiv.org/abs/2007.05084

[20] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526*. [Online]. Available: http://arxiv.org/abs/1712.05526

[21] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in CNNS by training set corruption without label poisoning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 101–105.

[22] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Proc. ECCV*, 2020, pp. 182–199.

[23] A. M. Nguyen and A. Tran, "WaNet–imperceptible warping-based backdoor attack," 2021, *arXiv:2102.10369*. [Online]. Available: https://arxiv.org/abs/2102.10369
[24] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu, and D. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation," in *Proc. 10th ACM Conf. Data Appl. Secur. Privacy*, Mar. 2020, pp. 1–15.
[25] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14431–14440.
[26] A. Chan, Y. Tay, Y.-S. Ong, and A. Zhang, "Poison attacks against text datasets with conditional adversarially regularized autoencoder," in *Proc. Findings Assoc. for Comput. Linguistics*, 2020, pp. 1–15.
[27] Z. Xi, R. Pang, S. Ji, and T. Wang, "Graph backdoor," 2020, *arXiv:2006.11890*. [Online]. Available: http://arxiv.org/abs/2006.11890
[28] A. Turner, D. Tsipras, and A. Madry, "Clean-label backdoor attacks," Tech. Rep., 2018.
[29] A. Shafahi, W. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Proc. NeurIPS*, 2018, pp. 1–18.
[30] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Proc. NeurIPS*, 2018, pp. 1–16.
[31] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," 2018, *arXiv:1811.03728*. [Online]. Available: http://arxiv.org/abs/1811.03728
[32] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 707–723.
[33] T. J. L. Tan and R. Shokri, "Bypassing backdoor detection algorithms in deep learning," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Sep. 2020, pp. 175–183.
[34] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," 2020, *arXiv:2003.03675*. [Online]. Available: http://arxiv.org/abs/2003.03675
[35] A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," 2020, *arXiv:2010.08138*. [Online]. Available: http://arxiv.org/abs/2010.08138
[36] J. Lin, L. Xu, Y. Liu, and X. Zhang, "Composite backdoor attack for deep neural network by mixing existing benign features," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 113–131.
[37] Y. Li, Y. Li, Y. Lv, Y. Jiang, and S.-T. Xia, "Hidden backdoor attack against semantic segmentation models," 2021, *arXiv:2103.04038*. [Online]. Available: http://arxiv.org/abs/2103.04038
[38] J. Dumford and W. Scheirer, "Backdooring convolutional neural networks via targeted weight perturbations," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–9.
[39] A. S. Rakin, Z. He, and D. Fan, "TBT: Targeted neural network attack with bit trojan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13195–13204.
[40] J. Bai, B. Wu, Y. Zhang, Y. Li, Z. Li, and S.-T. Xia, "Targeted attack against deep neural networks via flipping limited weight bits," 2021, *arXiv:2102.10496*. [Online]. Available: http://arxiv.org/abs/2102.10496
[41] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," 2021, *arXiv:2106.09667*. [Online]. Available: https://arxiv.org/abs/2106.09667
[42] D. Tang, X. Wang, H. Tang, and K. Zhang, "Demon in the variant: Statistical analysis of DNNs for robust backdoor contamination detection," 2019, *arXiv:1908.00686*. [Online]. Available: http://arxiv.org/abs/1908.00686
[43] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: A defence against trojan attacks on deep neural networks," in *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, Dec. 2019, pp. 113–125.
[44] T. Huster and E. Ekwedike, "TOP: Backdoor detection in neural networks via transferability of perturbation," 2021, *arXiv:2103.10274*. [Online]. Available: http://arxiv.org/abs/2103.10274
[45] Y. Zeng, W. Park, Z. Morley Mao, and R. Jia, "Rethinking the backdoor attacks' triggers: A frequency perspective," 2021, *arXiv:2104.03413*. [Online]. Available: http://arxiv.org/abs/2104.03413
[46] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. RAID*, 2018, pp. 273–294.
[47] Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, and S. Xia, "Rethinking the trigger of backdoor attack," 2020, *arXiv:2004.04692*. [Online]. Available: http://arxiv.org/abs/2004.04692
[48] J. Geiping, L. Fowl, G. Somepalli, M. Goldblum, M. Moeller, and T. Goldstein, "What doesn't kill you makes you robust (er): Adversarial training against poisons and backdoors," 2021, *arXiv:2102.13624*. [Online]. Available: http://arxiv.org/abs/2102.13624
[49] J. Carnerero-Cano, L. Mu noz-González, P. Spencer, and E. C. Lupu, "REGularization can helpmitigatepoisoningattacks…with therighthyperparameters," 2021, *arXiv:2105.10948*. [Online]. Available: https://arxiv.org/abs/2105.10948
[50] Y. Gao, B. Gia Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," 2020, *arXiv:2007.10760*. [Online]. Available: http://arxiv.org/abs/2007.10760
[51] A. Lagae, S. Lefebvre, R. L. Cook, T. DeRose, G. Drettakis, D. Ebert, J. P. Lewis, K. Perlin, and M. Zwicker, "State of the art in procedural noise functions," in *Proc. Eurographics*, 2010, pp. 1–19.
[52] K. Perlin, "An image synthesizer," *ACM SIGGRAPH Comput. Graph.*, vol. 19, no. 3, pp. 287–296, Jul. 1985.
[53] D. Mitrovic, "Real-time shading languages," Tech. Rep., 2015.
[54] A. Lagae, S. Lefebvre, G. Drettakis, and P. Dutré, "Procedural noise using sparse Gabor convolution," *ACM Trans. Graph.*, vol. 28, p. 54, Oct. 2009.
[55] S. Worley, "A cellular texture basis function," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 291–294.
[56] K. T. Co, L. Muñoz-González, S. de Maupeou, and E. C. Lupu, "Procedural noise adversarial examples for black-box attacks on deep convolutional networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 275–289.
[57] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 372–387.
[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019.
[59] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and surface variations," 2018, *arXiv:1807.01697*. [Online]. Available: http://arxiv.org/abs/1807.01697

XUAN CHEN received the B.S. degree from Air Force Engineering University, China, where he is currently pursuing the master's degree with the Department of Basic Sciences. His research interests include machine learning and adversarial attacks.

YUENA MA received the Ph.D. degree in electronic science and technology from Northwestern Polytechnical University. She is currently a Professor with Air Force Engineering University. Her research interests include self-orthogonal (or dual containing) code, quantum error-correcting code, communication security, neural networks, and deep learning.

SHIWEI LU received the B.S. degree in computer science and technology from Zhejiang University, in 2017, and the M.S. degree from Air Force Engineering University, China. He is currently a Doctor with the Department of Basic Sciences, Air Force Engineering University. His research interests include cyberspace security and machine learning.