A Dual Stealthy Backdoor: From Both Spatial and Frequency Perspectives

Yudong Gao[1], Honglong Chen[1*], Peng Sun[2], Junjian Li[1], Anqing Zhang[1], Zhibo Wang[3], Weifeng Liu[1]

[1]College of Control Science and Engineering, China University of Petroleum (East China), P.R. China
[2]College of Computer Science and Electronic Engineering, Hunan University, P.R. China
[3]School of Cyber Science and Technology, Zhejiang University, P.R. China
YudongGao0504@163.com, chenhl@upc.edu.cn, psun@hnu.edu.cn, ljj1016cc@163.com, zhanganqing@126.com, zhibowang@zju.edu.cn, liuwf@upc.edu.cn

Abstract

Backdoor attacks pose serious security threats to deep neural networks (DNNs). Backdoored models make arbitrarily (targeted) incorrect predictions on inputs containing well-designed triggers, while behaving normally on clean inputs. Prior researches have explored the invisibility of backdoor triggers to enhance attack stealthiness. However, most of them only focus on the invisibility in the spatial domain, neglecting the generation of invisible triggers in the frequency domain. This limitation renders the generated poisoned images easily detectable by recent defense methods. To address this issue, we propose a **DU**al stealthy **BA**ckdoor attack method named **DUBA**, which simultaneously considers the invisibility of triggers in both the spatial and frequency domains, to achieve desirable attack performance, while ensuring strong stealthiness. Specifically, we first use Wavelet Transform to embed the high-frequency information of the trigger image into the clean image to ensure attack effectiveness. Then, to attain strong stealthiness, we incorporate Fourier Transform and Cosine Transform to mix the poisoned image and clean image in the frequency domain. Moreover, DUBA adopts a novel attack strategy by training the model with weak triggers and attacking with strong triggers to further enhance attack performance and stealthiness. DUBA is evaluated extensively on four datasets against popular image classifiers, showing significant superiority over state-of-the-art backdoor attacks in attack success rate and stealthiness.

Introduction

Deep neural networks (DNNs) have made great achievements in many fields, such as image classification (He et al. 2016), image segmentation (Feng et al. 2022), and target recognition (Peng et al. 2021). Despite this remarkble success, DNNs are exposed to various security threats due to their reliance on datasets labeled by third parties or during outsourced training (Wang et al. 2023). Recent studies have shown the vulnerability of DNNs to backdoor attacks (Gu et al. 2017), where adversaries intentionally manipulate either the training data or model parameters to ensure accurate predictions on clean data while inducing targeted incorrect predictions on poisoned data. Backdoor attacks pose serious security threats to deep learning systems, especially in security-sensitive applications (e.g., autonomous driving). Most existing backdoor attacks craft backdoor triggers in the spatial domain (Gu et al. 2017; Chen et al. 2017; Li et al. 2021a). However, as shown in Figure 1, with the visible triggers, early backdoor attacks can be easily detected and removed. To enhance attack stealthiness, recent works consider generating invisible triggers through image steganography (Li et al. 2021a), disorting fields (Nguyen and Tran 2021) and so on. Regrettably, these methods only address the invisibility of triggers in the spatial domain but neglect that in the frequency domain. As a result, the generated backdoored images can be easily identified by typical image classification models that employ the frequency transform as a part of the task pipeline (Zhang et al. 2021; Yan et al. 2019). More importantly, researchers have proposed effective backdoor defenses from frequency perspective. For example, FTD (Zeng et al. 2021) demonstrated that most backdoor triggers are high-frequency semantics, and it trained a DNN model that classifies images in the frequency domain to effectively defend against most attacks. That is, most backdoors are perceptible in the frequency domain. Therefore, it becomes imperative to guarantee the trigger's imperceptibility in both the spatial and frequency domains to launch a powerful yet stealthy backdoor attack.

To the best of our current understanding, the scholarly dis-

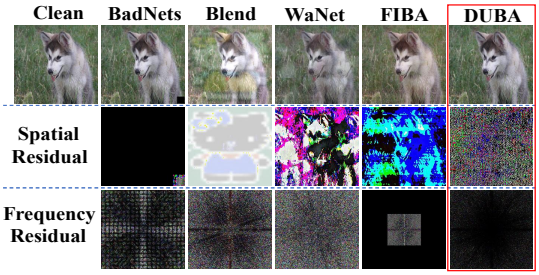

Figure 1: Visualization of images poisoned by different attacks. The first row contains: the clean image, images poisoned by BadNets (Gu et al. 2017), Blend (Chen et al. 2017), WaNet (Nguyen and Tran 2021), FIBA (Feng et al. 2022), and the proposed DUBA. The second row is the corresponding residual images in the spatial domain and the bottom is the residual images in the frequency domain.

*Corresponding author.

cussion concerning the imperceptibility of triggers from the frequency perspective seems to be significantly limited. For instance, the study in (Zeng et al. 2021) utilizes a low-pass filter to embed a backdoor that remains invisible in the frequency domain but becomes apparent in the spatial domain.

Motivated by the above discussions, in this paper, we propose a **DU**al stealthy **BA**ckdoor attack called **DUBA**, which crafts invisible triggers in both the spatial and frequency domains while achieving desirable attack performance. Specifically, we first embed the high-frequency information of the trigger image into the clean image by discrete wavelet transform (DWT), yielding the initial poisoned image. Then, to ensure strong stealthiness, we fuse the initial poisoned image (which has high-frequency triggers) with the clean image in the Fourier and Cosine transform domains. Furthermore, we propose an attack strategy, which reduces the embedded high-frequency information of the trigger images and expands the masking ratio of the trigger images in the training phase to make the victim model better learn the triggers. The major contributions of this paper are summarized as follows:

- We design a **DU**al stealthy **BA**ckdoor attack named DUBA that achieves desirable invisibility in both spatial and frequency domains by embedding high-frequency trigger information through DWT and smoothing it in Fourier Transform and Cosine Transform domains.
- We propose a novel attack strategy for DUBA where the model is trained with weak triggers and attacked with strong triggers to attain satisfactory attack performance while ensuring stealthiness.
- We conduct extensive experimental evaluation of DUBA on four datasets and popular models. The results demonstrate its outstanding performance in terms of both attack effectiveness and stealthiness.

Related Work

Backdoor Attack

Backdoor attack has drawn wide attention since its introduction. According to the trigger generation method, existing backdoors can be roughly divided into two categories, i.e., spatial domain backdoors and frequency domain backdoors.

Spatial Domain Backdoors BadNets (Gu et al. 2017) first reveals the existence of backdoors in DNNs. This attack embeds a visible square in the bottom right corner of the clean image and manipulates the associated label to the target label. Then, the backdoor can be injected to the model after training on the poisoned data. In the inference phase, images with the same trigger will be misclassified into the attacker-chosen label. Inspired by BadNets, researchers have also investigated other backdoor attacks. Blend (Chen et al. 2017) advocates image blending backdoors, whereas another work (Jacob, Pang, and Percy 2017) employs a fixed watermark as a trigger to insert backdoors. However, these early backdoors are visually visible, making them easily detected and removed. Therefore, how to generate visually invisible backdoors has recently become a hot research topic. For example, ISSBA (Li et al. 2021a) embeds trigger information by steganography; WaNet (Nguyen and Tran 2021) crafts triggers by distorting fields; and LIRA (Doan et al. 2021) searches for triggers in a highly nonlinear parameter space. While these methods generate invisible triggers and evade mainstream backdoor defenses, they neglect image invisibility in the frequency domain. Consequently, models empowered by Fourier transform (commonly employed in the task pipeline) or frequency-oriented defense methods can easily detect these backdoor attacks.

Frequency Domain Backdoors Recently, (Zeng et al. 2021) starts to explore backdoor attacks in the frequency domain. To avoid high-frequency artifacts after the Discrete Cosine Transform (DCT) (Cintra and Bayer 2011), a low-pass filter is applied to generate a smooth trigger. However, this method yields visible artifacts in the spatial domain. FIBA (Feng et al. 2022) crafts triggers in the frequency domain by mixing the low-frequency components of two images after Fast Fourier Transform (FFT) (Moreland and Angel 2003), which is visually imperceptible in spatial domain but still visible in frequency domain. Similarly, another work, FTROJAN (Wang et al. 2022), transforms the clean image using YUV or UV color coding methods and then applies DCT with modifications to the high-frequency or mid-frequency components to generate the poisoned image. However, the trigger generated in FTROJAN is also visible in the frequency domain.

Backdoor Defense

To defend various backdoor attacks, researchers have proposed many defense methods (Li et al. 2021b; Xu et al. 2020). Generally, backdoor defenses can be categorized into input-based, model-based, and output-based methods.

Input-based Defenses Input-based defenses focus on input abnormalities (Chou, Tramer, and Pellegrino 2020; Zeng et al. 2021). Grad-Cam (Selvaraju et al. 2017) uses a saliency map to dissect the regions of the input image that the model focuses on. If the model does not focus on the object or keeps focusing on the same region, the image is considered as poisoned. FTD (Zeng et al. 2021) employs DCT to distinguish whether the input image has high-frequency artifacts. They design a DNN-based discriminator to classify images with high-frequency artifacts as poisoned images.

Model-based Defenses Their emphasis lies in investigating the victim model (Kolouri et al. 2020). Fine-Pruning (Liu, Dolan-Gavitt, and Garg 2018) combats backdoors by pruning dormant neurons, as these neurons likely offer specialized support to backdoors. Neural Cleanse (Wang et al. 2019) identifies backdoors in the model through trigger reverse engineering and leverages anomaly detection to pinpoint the most potential backdoor.

Output-based Defenses Defense methods of this type often observe output anomalies (Huang et al. 2020; Gao et al. 2019). STRIP (Gao et al. 2019) superimposes various image patterns on the suspicious image to observe its output. Higher poisoning odds yield lower output randomness. To evade the existing backdoor defenses from both spatial and frequency domains, in this work, we aim to craft a powerful backdoor attack that is invisible in both domains.

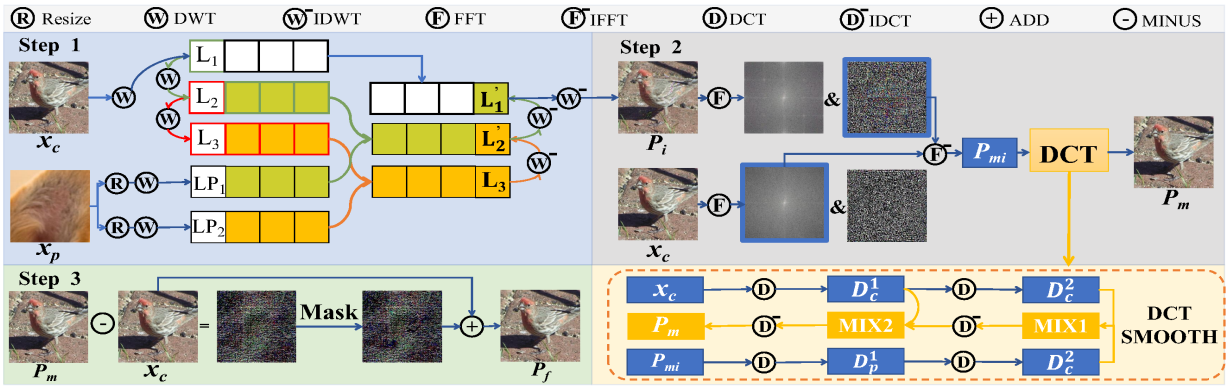

Figure 2: Schematic diagram of DUBA. Step 1: high-frequency information embedding. Embed the high-frequency information of the trigger image into the clean image by DWT. Note that the initial trigger x_p is a randomly selected image. Step 2: frequency domain smoothing. Smooth the high-frequency trigger in the FFT and DCT domains. Step 3: random trigger masking. Mask the trigger embedded within image P_m when the corresponding clean image pixel values are close to 0 or 255.

Methodology

Threat Model

Attacker's Capabilities In the training phase, following prior studies (Li et al. 2021a; Zhong, Qian, and Zhang 2022), the attacker can only tamper with a part of the training data, lacking access to other model training components (e.g., victim model architecture and loss function). In the inference phase, the attacker can only manipulate the input images (i.e., embedding the crafted backdoor trigger). This threat model is common in real-world scenarios, such as outsourcing model training to third parties.

Attacker's Goals Generally, an effective backdoor attack should deceive models into making arbitrarily (targeted) incorrect predictions on tampered testing images, while maintaining normal performance on clean inputs. Moreover, a potent backdoor attack must satisfy two objectives: **invisibility**, ensuring the poisoned images remain indistinguishable in both spatial and frequency domains, and **robustness**, enabling it to bypass state-of-the-art defense methods.

Problem Formulation

We focus on supervised image classification, a widely used technique in face recognition, traffic signal recognition, and other security-sensitive applications. Formally, the image classification can be described as a mapping function $f_\theta : \mathcal{X} \to \mathcal{C}$, where $\mathcal{X}$ is the input domain and $\mathcal{C}$ is the set of target classes. The model parameters θ are learned from the training dataset $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ of N data samples, where $x_i \in \mathcal{X}$, $y_i \in \mathcal{C}$. The core of backdoor attacks is to craft a set of posioned data samples $D_{\text{poison}} = \{(T(x_i), \gamma(y_i))\}_{i=1}^{M}$, where $T(\cdot)$ denotes the trigger implantation method and $\gamma(\cdot) \in \mathcal{C}$ represents the designated target label. Specifically, $\gamma(y_i) = c$, where c is a constant, stands for the All-to-One attack, while $\gamma(y_i) = y_i + 1$ represents All-to-All attack. In summary, backdoor attacks aim to manipulate a subset of the training data (note that $M < N$ and $\partial = \frac{M}{N}$ is the poisoning ratio) by injecting adversarial triggers such that the model trained on the tampered dataset yields the following behaviors when deployed:

$$f_\theta(x_i) = y_i, \quad f_\theta(T(x_i)) = \gamma(y_i). \tag{1}$$

This paper is centered on the design of the trigger implantation method $T(\cdot)$.

The Proposed Attack

Overview of DUBA Figure 2 shows the framework of DUBA, which is composed of three steps and an attack strategy. First, to attain desirable backdoor attack performance, we employ DWT to embed the high-frequency information of a fixed trigger image into the clean image to generate the initial poisoned image. Second, to ensure strong stealthiness in both spatial and frequency domains, we incorporate FFT and DCT to mix the initial poisoned image with clean image to generate the intermediate poisoned image. Third, to ensure that the victim model learns the backdoor scattered over the entire poisoned image while maintaining good invisibility, we propose to randomly mask the trigger of intermediate poisoned image to get the final poisoned image. Moreover, we propose an attack strategy where the victim model is trained with weak triggers and attacked with strong triggers to achieve higher attack success rates (ASRs).

Step 1: High-Frequency Information Embedding Inspired by prior works (Zeng et al. 2021) that utilize high-frequency semantic information for trigger embedding, we propose to extract the high-frequency information from a fixed image as the initial trigger. Specifically, we employ DWT for high-frequency information extraction given that DWT can finely dissect images at high frequencies but coarsely analyze images at low frequencies. In Step 1 of Figure 2, given a clean image x_c and a random initial trigger image x_p, the idea is to embed the high-frequency part of x_p into the deep high-frequency region of x_c. DWT decomposes image x into one low-frequency part and three high-frequency parts, represented as follows:

$$W(x) = \{L, H_1, H_2, H_3\}, \tag{2}$$

where L represents the low-frequency approximate component, while H_1, H_2, and H_3 denote the high-frequency components at the vertical, diagonal, and horizontal directions, respectively. Accordingly, the image can also be recovered by inverse discrete wavelet transform (IDWT) W^-:

$$W^- \left(L, H_1, H_2, H_3\right) = x. \tag{3}$$

To embed a sufficiently hidden trigger, we apply three DWTs to clean image x_c, which is expressed as:

$$\left\{L_{i+1}, H_{i+1,1}, H_{i+1,2}, H_{i+1,3}\right\} = W\left(L_i\right), i = 0, 1, 2, \tag{4}$$

where L_0 equals x_c and i stands for the i-th DWT. Then, we apply one DWT to trigger image x_p. Since the image size changes after three DWTs on x_c, trigger image x_p needs to be resized into two different sizes, denoted as x_{p1} and x_{p2}. For both x_{p1} and x_{p2}, a single DWT is applied to obtain two distinct high-frequency trigger information:

$$\begin{array}{l} \left\{LP_1, HP_{1,1}, HP_{1,2}, HP_{1,3}\right\} = W(x_{p1}) \\ \left\{LP_1^{'}, HP_{1,1}^{'}, HP_{1,2}^{'}, HP_{1,3}^{'}\right\} = W(x_{p2}) \end{array}. \tag{5}$$

Next we embed the high-frequency information of the trigger image, $HP_{1,j}$ and $HP_{1,j}^{'}$ (j = 1, 2, 3), into different high-frequency parts of the clean image x_c. Formally, we have:

$$\left\{\begin{array}{l} H_{3,j}^{'} = H_{3,j} \times \alpha + HP_{1,j}^{'} \times (1-\alpha) \\ H_{2,j}^{'} = H_{2,j} \times \beta + HP_{1,j} \times (1-\beta) \end{array}\right., \ j = 1, 2, 3, \tag{6}$$

where α and β indicate the embedding intensity. With $H_{3,j}^{'}$ and $H_{2,j}^{'}$, initial poisoned image P_i can be derived using the IDWT as follows:

$$L_i^{'} = W^- \left(L_{i+1}^{'}, H_{i+1,1}^{'}, H_{i+1,2}^{'}, H_{i+1,3}^{'}\right), i = 0, 1, 2, \tag{7}$$

where $L_3^{'}$ equals L_3, $H_{1,j}^{'}$ equals $H_{1,j}$ and the final low-frequency component $L_0^{'}$ is exactly the generated poisoned image in Step 1, which is denoted as P_i. The above method generates an almost invisible backdoor image in the spatial domain due to the stealthiness of high-frequency information (subject to the intensity of α and β, discussed in ablation experiments). In what follows, we aim to answer the question on how to craft a poisoned image that is also invisible in the frequency domain.

Step 2: Frequency Domain Smoothing Previous studies (Govindaraju et al. 2008) have highlighted the significance of the phase spectrum in retaining high-level semantic information, including edges and overall image structure, after FFT. Meanwhile, the amplitude spectrum captures underlying frequency information (Feng et al. 2022). Research (Yang and Soatto 2020) further indicates that changes in the amplitude spectrum have limited impact on the perception of high-level semantics. Considering the complex domain output of FFT and the amplitude spectrum is usually observed in the frequency domain, we adopt a straightforward yet effective approach. Specifically, to ensure both the backdoor attack performance and the amplitude spectrum's invisibility, we directly swap the amplitude spectrums of the P_i and x_c after FFT. Formally, as shown in Step 2 of Figure 2, let $F^S\left(\cdot\right)$ and $F^P\left(\cdot\right)$ be the amplitude and phase components of FFT results, the amplitude and phase spectra of P_i and x_c after FFT are obtained as:

$$\left\{\begin{array}{l} S^c = F^S\left(x_c\right), P^c = F^P\left(x_c\right) \\ S^p = F^S\left(P_i\right), P^p = F^P\left(P_i\right) \end{array}\right. . \tag{8}$$

Then the smoothing poisoned image P_{mi} is calculated by the amplitude spectrum of x_c and the phase spectrum of P_i:

$$P_{mi} = F^-\left(S^c, P^p\right), \tag{9}$$

where $F^-(\cdot)$ represents the inverse FFT. Note that the FFT-based smoothing is conducted in the complex domain. Thus, despite the spectrogram of the image being theoretically hidden after FFT-based smoothing, the image is still perceivable in the real domain (see the ablation experiment for a visual demonstration). Moreover, the poisoned image may be detected by the DCT-based defense, which is unacceptable even though the detection probability is low.

To address this issue, next we incorporate the DCT, which is a special case of FFT in real domain, to fuse P_{mi} and x_c. Due to the linear property of DCT, fusing the two images after only one DCT and then inversing the fused result is equivalent to directly fusing the two images, which can not achieve the purpose of deep smoothing. Thus we apply two DCTs on P_{mi} and x_c to achieve deeper information fusion. Let D be the DCT while D^- be the inverse DCT (IDCT), the deep information in DCT domain is obtained as follows:

$$D_c^k = D\left(D_c^{k-1}\right), D_p^k = D\left(D_p^{k-1}\right), \quad k = 1, 2, \tag{10}$$

where k stands for the k-th DCT, D_c^0 equals x_c, and D_p^0 equals P_{mi}. As shown in Figure 2, the DCT smoothing is then implemented according to:

$$D_p^{k-1} = D^-\left[D_p^k \times \lambda + D_c^k \times (1-\lambda)\right], \quad k = 1, 2, \tag{11}$$

where λ indicates the fusing intensity. After two steps of IDCT, we obtain the new D_p^0, which is exactly the intermediate poisoned image (denoted as P_m) generated in Step 2.

Step 3: Random Trigger Masking To ensure that the victim model learns the backdoor scattered over the entire image while well preserving the desirable attack stealthiness in both the spatial and frequency domains, we propose to randomly mask the trigger image. As shown in Step 3 of Figure 2, we first obtain the trigger embedded in P_m by subtracting x_c from P_m. Then, randomly mask it. Finally, we again embed the trigger pattern after masking into the clean image, yielding the final poisoned image P_f.

Attack Strategy Design We further devise an attack strategy for DUBA to enhance both attack performance and stealthiness. Specifically, in the training phase, we adopt a weak trigger pattern via two operations, i.e., minimizing the values of α and β and masking more pixel points of the trigger image in Step 3. In the inference phase, both α and β are moderately amplified while ensuring the triggers' invisibilities and masking fewer pixel points in Step 3. For masking design, we consider triggers becoming visible where clean image pixel values are near 0 or 255. To address this, we mask corresponding regions in the trigger embedded in P_m.

Models	DataSet→ / Methods↓	Cifar10		Gtsrb		ImageNet		Fer2013	
		BA (%)	ASR (%)	BA (%)	ASR (%)	BA (%)	ASR (%)	BA (%)	ASR (%)
ResNet18	Clean	91.91		99.32		92.12		63.22	
	BadNets	91.22	99.36	99.14	99.62	91.56	99.17	62.72	**99.99**
	Blend	91.35	99.97	99.06	99.72	91.28	98.11	63.09	99.72
	WaNet	91.25	99.78	99.07	99.81	91.62	99.13	62.41	99.01
	FIBA	91.08	99.26	99.22	98.91	92.02	98.96	63.08	99.82
	DUBA	91.55	**99.98**	99.21	**99.92**	91.55	**99.24**	62.65	99.89
RepVGG	Clean	91.23		99.46		93.25		61.59	
	BadNets	91.08	99.34	99.42	99.87	92.88	**99.29**	61.52	99.88
	Blend	91.11	95.74	99.13	99.56	92.76	97.02	60.55	99.90
	WaNet	91.11	99.51	99.17	99.28	92.36	99.17	61.28	99.72
	FIBA	90.08	99.11	99.29	99.16	92.72	98.94	61.50	99.86
	DUBA	91.18	**99.98**	99.41	**99.89**	92.52	99.02	60.99	**99.91**
Conformer	Clean	92.92		99.50		93.28		63.88	
	BadNets	92.75	**99.24**	99.39	99.49	92.82	**99.27**	63.70	99.89
	Blend	92.51	98.92	99.45	99.06	93.18	98.27	63.72	99.84
	WaNet	92.36	99.03	99.26	98.56	93.06	99.01	63.62	99.75
	FIBA	92.08	98.93	99.27	98.69	93.18	98.81	62.51	99.82
	DUBA	92.44	99.12	99.39	**99.57**	93.11	99.19	63.21	**99.92**

Table 1: Attack effectiveness in terms of BA (%) and ASR (%).

Experiments

Experimental Settings

Datasets & Models We evaluate DUBA's performance on four datasets: Cifar10 (Krizhevsky, Hinton et al. 2009), Gtsrb (Stallkamp et al. 2012), ImageNet (Deng et al. 2009), and Fer2013 (Goodfellow et al. 2013). Note that the ImageNet is too large so we use a subset of it (100 classes). Three models are used to verify DUBA: ResNet18 (He et al. 2016), RepVGG (Ding et al. 2021), and Conformer (Peng et al. 2021).

Baseline Backdoor Attacks We compare DUBA with BadNets (Gu et al. 2017), Blend (Chen et al. 2017), WaNet (Nguyen and Tran 2021), and FIBA (Feng et al. 2022). BadNets and Blend are representational visible backdoor attacks. WaNet is the latest invisible backdoor attack in the spatial domain while FIBA is the latest backdoor attack proposed from the frequency perspective.

Evaluation Metrics We assess DUBA from two perspectives: attack performance and attack stealthiness. For attack performance evaluation, we employ the attack success rate (ASR), which measures the ratio of poisoned examples misclassified as the target label to all poisoned examples used for testing. Additionally, we utilize the benign accuracy (BA) to characterize the model's performance on clean testing data. For attack stealthiness evaluation, we use the following similarity metrics: peak signal-to-noise ratio (PSNR) (Tanchenko 2014), structural similarity (SSIM) (Hore and Ziou 2010) and learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018).

Implementation Details In our experiments, we randomly select an image with a dog's ear as the initial trigger. During training, α and β are set to 0.4. In the attack phase, α and β are both set to 0.6, λ to 0.7. All attacks are All-to-One as in other studies (Li et al. 2021a; Feng et al. 2022). Defense experiments are conducted on the RepVgg model. Further details can be found in the Appendix.

Attack Performance Evaluation

Attack Effectiveness We evaluate the effectiveness of different backdoor attacks with ASR and BA. The relevant results are summarized in Table 1, which shows that our proposed DUBA achieves higher or comparable ASRs under most datasets and models. But in some cases such as experiments under ImageNet, the ASRs of DUBA are slightly lower than BadNets. Considering that the crafted trigger by DUBA is invisible in both spatial and frequency domains (which will be validated), such a result is acceptable. Besides, DUBA only incurs negligible loss (lower than 1%) of BA compared with the clean benchmark. The above results show that DUBA achieves desirable attack effectiveness.

Attack Stealthiness Figure 1 shows the poisoned images of different methods and more visual comparison between clean images and images are in the Appendix. Compared with other methods, DUBA achieves the best invisibility in both the spatial and frequency domains. The backdoor generated by DUBA is visually invisible in the spatial domain and the residual image in the frequency domain is also close to pure black image, indicating its high similarity to the clean image in the frequency domain. Table 2 quantifies the visual outcomes, with all three metrics of DUBA being the best in most cases. Although DUBA's SSIM is slightly lower than BadNets, it remains close to 1 and higher than most methods. BadNets, with its obvious trigger, exhibits the worst stealthiness, as evident in Figure 1. Overall, DUBA achieves the best stealthy results.

Robustness to Defenses

In this subsection, we test DUBA against five state-of-the-art defenses, including GradCam (Selvaraju et al. 2017),

Methods	BadNets			Blend			WaNet			FIBA			DUBA		
METRIC	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Cifar10	30.19	0.970	0.0096	18.96	0.862	0.1219	33.98	0.966	0.0090	29.46	0.962	0.0115	**37.98**	**0.977**	**0.0081**
Gtsrb	34.51	**0.972**	0.0351	19.81	0.887	0.1661	30.51	0.960	0.0588	30.17	0.969	0.0622	**36.01**	**0.972**	**0.0201**
ImageNet	35.14	**0.979**	0.0079	20.56	0.896	0.1831	29.65	0.951	0.0932	30.56	0.969	0.0764	**36.22**	0.976	**0.0072**
Fer2013	30.28	**0.984**	0.0087	16.56	0.824	0.1957	31.26	0.962	**0.0079**	35.46	0.983	0.0083	**36.14**	0.982	0.0081

Table 2: Stealthiness of different attacks in the spatial domain.

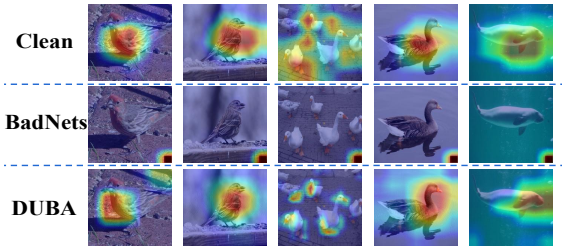

Figure 3: The GradCam of different samples. GradCam effectively defends against BadNets. For DUBA, it resembles clean images, with even stronger object locking.

Methods	Clean	BadNets	Blend	WaNet	FIBA	DUBA
Index	0.76	4.28	3.26	2.32	1.92	**1.22**

Table 3: Anomaly index values for different attacks.

Neural Cleanse (Wang et al. 2019), STRIP (Gao et al. 2019), Fine-Prunning (Liu, Dolan-Gavitt, and Garg 2018), and FTD (Zeng et al. 2021), as well as other transforms.

Robustness to GradCam GradCam calculates heat values for input images. For clean images, it focuses on the object. However, as shown in Figure 3, when dealing with small triggers like BadNets, the heat map concentrates on the trigger, leading to an abnormal heat map. For DUBA, GradCam results are similar to clean images and even highlight the object more distinctly. This suggests that GradCam fails to detect DUBA.

Robustness to Neural Cleanse Neural Cleanse reconstructs triggers for each label and checks for significantly small reverse-engineered triggers. Specifically, the method uses the anomaly index to quantify deviations based on trigger sizes, considering models with an anomaly index greater than 2 as poisoned. Table 3 shows DUBA's anomaly index at only 1.22, smaller than baseline methods. This confirms DUBA's effectiveness in evading Neural Cleanse.

Robustness to STRIP STRIP identifies model poisoning by overlaying input images and assessing prediction consistency through entropy values. Specifically, poisoned models have an average entropy value below 0.2 and inconsistency with clean results. Figure 4 shows entropy values for various methods on Cifar10. DUBA's entropy values surpass 0.2, closely aligning with clean results, outperforming BadNets and Blend significantly. DUBA also achieves comparable entropy results to WaNet and FIBA. Thus, DUBA can effectively bypass STRIP.

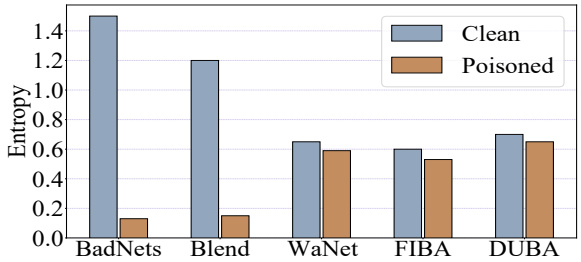

Figure 4: Entropy values of STRIP on Cifar10. The entropy values of poisoned images on DUBA closely resemble those of clean images.

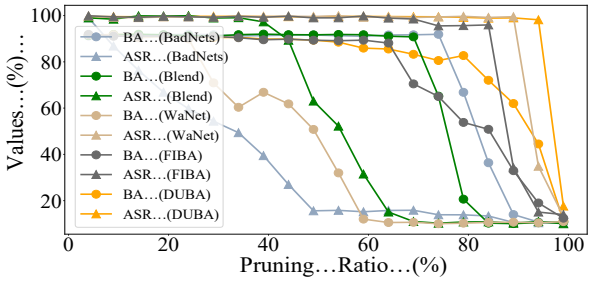

Figure 5: The results of Fine-Pruning on different methods. The ASR of DUBA is the latest to start declining.

Robustness to Fine-Pruning Fine-pruning links backdoor behavior to dormant neurons. Specifically, the neurons are pruned according to neuronal activation values. Usually, the attack is considered successful if the BA of clean images drops below 50% before the ASR of poisoned images. Figure 5 displays Cifar10 results for different methods. Despite slightly lower BA than WaNet, DUBA's ASR decline is slowest, reaching nearly 96% pruning ratio, and BA drops below 50% before ASR. Thus DUBA effectively counters pruning-based defenses.

Robustness to FTD FTD detects high-frequency artifacts in images as indicators of poisoned images. It employs a DNN model to classify images after DCT as poisoned or clean. Table 4 shows FTD's detection rate for DUBA is below 50%, suggesting effective evasion. This is due to DUBA's double smoothing of trigger images in the frequency domain. FIBA also achieves a low detection ratio

Methods	BadNets	Blend	WaNet	FIBA	DUBA
Rate (%)	91.72	96.41	68.95	58.2	**49.96**

Table 4: Detection rates of FTD for different attack methods.

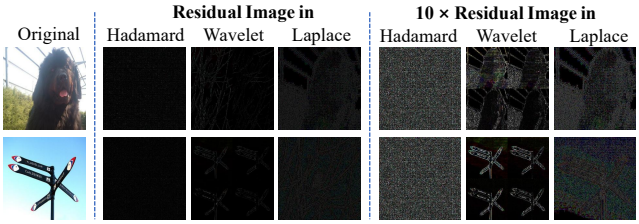

Figure 6: Visual results using other transforms. Left: Two original images. Middle: The residual images in Hadamard, Wavelet and Laplace frequency domains. Right: The corresponding residual images enlarged tenfold.

(higher than DUBA) due to its low-frequency trigger.

Robustness to Other Frequency Transforms We investigated trigger invisibility in the frequency domain through stealth experiments using transforms beyond FFT and DCT. Figure 6 indicates triggers largely stay hidden in these transforms. Visible perturbations emerge only after about tenfold amplification. We propose that strong correlations among various frequency domain transforms may account for this phenomenon. Consequently, we confidently predict that the suggested triggers are highly probable to elude future defenses in alternative frequency domains.

Ablation Studies

High-Frequency Embedding Rate We first examine the effect of α and β (trigger image embedding ratios) on ASR. Table 5 shows that higher embedding ratios during training and attacking phases lead to increased ASRs with DUBA, aligning with our expectations. Notably, practical considerations require careful attention to the trigger's concealability.

DCT Smoothing Parameters We further study the DCT smoothing parameter's impact on ASR. Intuitively, as λ decreases, poisoned images become closer to clean ones, compromising ASR but enhancing stealthiness. Table 6 validates our intuition, showing an ASR increase with decreasing λ.

Initial Trigger Selection We investigate the impact of initial trigger images on DUBA. Apart from the dog's ear image used in the previous experiments, three other images from Cifar10, Gtsrb, and ImageNet datasets are also tested. Table 7 indicates that the initial trigger has no significant correlation with ASR.

Train (α & β)	0.2	0.2	0.4	0.4	0.6
Attack (α & β)	0.2	0.6	0.6	0.8	0.6
ASR (%)	85.62	87.92	99.98	99.99	99.22

Table 5: Impact of high-frequency embedding ratio on ASR.

λ	0.1	0.2	0.3	0.5	0.7	0.8
ASR(%)	20	52.36	70.11	95.42	99.98	99.98

Table 6: Impact of DCT smoothing parameter λ on ASR.

Trigger Image	Dog's ear	Cifar10	Gtsrb	ImageNet
ASR (%)	99.98	99.23	99.57	99.62

Table 7: Impact of different initial trigger images on ASR.

The Necessity of Three Frequency Transforms We show the visual results using different transforms. As depicted in Figure 7, employing only DWT leads to visible artifacts in the frequency domain of the poisoned images. Alternatively, using FFT for smoothing renders the trigger invisible in the frequency domain, yet some artifacts persist in the spatial domain (indicated by red circles). By incorporating all three transforms, the poisoned image becomes stealthy in both domains, resulting in improved PSNR and SSIM values. Thus the proposed DUBA adopts these three frequency domain transforms to achieve dual stealth.

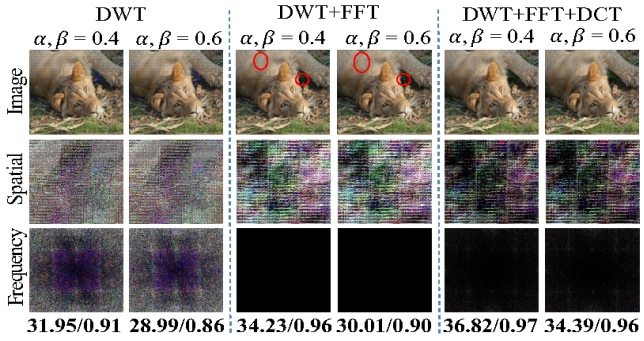

Figure 7: Visual results using different combinations of transforms. Poisoned images with different transforms (first row). Corresponding residual images in the spatial (second row) and frequency domains (third row). PSNR/SSIM values for each image (bottom).

Conclusion

In this paper, we showed that most backdoor attacks are visible in the frequency domain. In order to completely break the defense proposed from the frequency perspective while remaining stealthy in the spatial domain, we proposed a **DU**al stealthy **BA**ckdoor called DUBA that is invisible in both the spatial and frequency domains. To hide high-frequency backdoor information in both the spatial and frequency domains, we leveraged the benefits of different frequency domain transforms. A novel attack strategy was also devised in order to enhance the efficiency of DUBA. We conducted an extensive experimental evaluation of DUBA. The results corroborate its outstanding performance in terms of attack success rates and attack stealthiness. [1]

[1]codes: https://github.com/ifen1/Dual-Stealthy-Backdoor

Acknowledgments

This work was supported in part by the Shandong Provincial Natural Science Foundation, China, under Grant ZR2022YQ61, NSFC under Grants 61772551, 62111530052, 62102337, 62122066, the Shandong Provincial Natural Science Foundation, China, under Grants ZR2023ZD32, ZR2023MF008, the Natural Science Foundation of Hunan Province of China under Grant 2023JJ40174, the Fundamental Research Funds for the Central Universities, China, under Grant 22CX01003A-9.

References

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

Chou, E.; Tramer, F.; and Pellegrino, G. 2020. Sentinet: Detecting localized universal attacks against deep learning systems. In *IEEE Security and Privacy Workshops*, 48–54.

Cintra, R. J.; and Bayer, F. M. 2011. A DCT approximation for image compression. *IEEE Signal Processing Letters*, 579–582.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; and Sun, J. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13733–13742.

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

Feng, Y.; Ma, B.; Zhang, J.; Zhao, S.; Xia, Y.; and Tao, D. 2022. FIBA: Frequency-Injection based Backdoor Attack in Medical Image Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20876–20885.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 113–125.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *Proceedings of the International Conference on Neural Information Processing*, 117–124.

Govindaraju, N. K.; Lloyd, B.; Dotsenko, Y.; Smith, B.; and Manferdelli, J. 2008. High performance discrete Fourier transforms on graphics processors. In *Proceedings of the ACM/IEEE conference on Supercomputing*, 1–12.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *IEEE Access*, 47230–47244.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *Proceedings of International Conference on Pattern Recognition*, 2366–2369.

Huang, S.; Peng, W.; Jia, Z.; and Tu, Z. 2020. One-pixel signature: Characterizing cnn models for backdoor detection. In *European Conference on Computer Vision*, 326–341.

Jacob, S.; Pang, W. K.; and Percy, L. 2017. Certified defenses for data poisoning attacks. In *Proceedings of the International Conference on Neural Information Processing Systems*, 3520–3532.

Kolouri, S.; Saha, A.; Pirsiavash, H.; and Hoffmann, H. 2020. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 301–310.

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

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021a. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16463–16472.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021b. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, 273–294.

Moreland, K.; and Angel, E. 2003. The FFT on a GPU. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware*, 112–119.

Nguyen, A.; and Tran, A. 2021. WaNet: Imperceptible Warping-based Backdoor Attack. *arXiv preprint arXiv:2102.10369*.

Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; and Ye, Q. 2021. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 367–376.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 618–626.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32: 323–332.

Tanchenko, A. 2014. Visual-PSNR measure of image quality. *Journal of Visual Communication and Image Representation*, 874–878.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and

mitigating backdoor attacks in neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy*, 707–723.

Wang, T.; Yao, Y.; Xu, F.; An, S.; Tong, H.; and Wang, T. 2022. An Invisible Black-Box Backdoor Attack Through Frequency Domain. In *Proceedings of the European Conference on Computer Vision*, 396–413.

Wang, X.; Chen, H.; Sun, P.; Li, J.; Zhang, A.; Liu, W.; and Jiang, N. 2023. AdvST: Generating Unrestricted Adversarial Images via Style Transfer. *IEEE Transactions on Multimedia*, 1–13.

Xu, K.; Liu, S.; Chen, P.-Y.; Zhao, P.; and Lin, X. 2020. Defending against backdoor attack on deep neural networks. *arXiv preprint arXiv:2002.12162*.

Yan, T.; Wu, J.; Zhou, T.; Xie, H.; Xu, F.; Fan, J.; Fang, L.; Lin, X.; and Dai, Q. 2019. Fourier-space diffractive deep neural network. *Physical Review Letters*, 023901.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4085–4095.

Zeng, Y.; Park, W.; Mao, Z. M.; and Jia, R. 2021. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16473–16481.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, X.; Cui, P.; Xu, R.; Zhou, L.; He, Y.; and Shen, Z. 2021. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5372–5382.

Zhong, N.; Qian, Z.; and Zhang, X. 2022. Imperceptible Backdoor Attack: From Input Space to Feature Representation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1736–1742.