FIBA: Frequency-Injection based Backdoor Attack in Medical Image Analysis

Yu Feng[1*] Benteng Ma[1*] Jing Zhang[2] Shanshan Zhao[3] Yong Xia[1†] Dacheng Tao[3,2]
[1] School of Computer Science and Engineering, Northwestern Polytechnical University, China
[2] The University of Sydney, Australia, [3] JD Explore Academy, China
{fengy,mabenteng} @mail.nwpu.edu.cn,jing.zhang1@sydney.edu.au
sshan.zhao00@gmail.com, yxia@nwpu.edu.cn,dacheng.tao@gmail.com

Abstract

*In recent years, the security of AI systems has drawn increasing research attention, especially in the medical imaging realm. To develop a secure medical image analysis (MIA) system, it is a must to study possible backdoor attacks (BAs), which can embed hidden malicious behaviors into the system. However, designing a unified BA method that can be applied to various MIA systems is challenging due to the diversity of imaging modalities (*e.g.*, X-Ray, CT, and MRI) and analysis tasks (*e.g.*, classification, detection, and segmentation). Most existing BA methods are designed to attack natural image classification models, which apply spatial triggers to training images and inevitably corrupt the semantics of poisoned pixels, leading to the failures of attacking dense prediction models. To address this issue, we propose a novel Frequency-Injection based Backdoor Attack method (FIBA) that is capable of delivering attacks in various MIA tasks. Specifically, FIBA leverages a trigger function in the frequency domain that can inject the low-frequency information of a trigger image into the poisoned image by linearly combining the spectral amplitude of both images. Since it preserves the semantics of the poisoned image pixels, FIBA can perform attacks on both classification and dense prediction models. Experiments on three benchmarks in MIA (*i.e.*, ISIC-2019 [4] for skin lesion classification, KiTS-19 [17] for kidney tumor segmentation, and EAD-2019 [1] for endoscopic artifact detection), validate the effectiveness of FIBA and its superiority over state-of-the-art methods in attacking MIA models and bypassing backdoor defense. Source code will be available at code.*

1. Introduction

Deep neural networks (DNNs) are increasingly deployed in computer-aided diagnosis (CAD) systems and have achieved diagnostic parity with medical professionals on radiology, pathology, dermatology, and ophthalmology tasks [52]. However, recent studies have shown that DNNs are vulnerable to various attacks during the model's training and inference [8, 14, 23, 38]. Typically, attacks in the inference stage take the form of the adversarial samples [10, 45] and attempt to fool a trained model by manipulating the input. Backdoor attacks, in contrast, seek to maliciously alter the model in the training phase [3, 11, 35]. Although the research on adversarial samples has experienced rapid development recently, backdoor attacks have received less attention, especially in medical image analysis (MIA).

In general, backdoor attacks aim to embed a hidden backdoor trigger into DNNs so that the injected model performs well on benign testing samples when the backdoor is not activated, however, once the backdoor is activated by the attacker, the prediction will be changed to the target label as attackers expected [3, 11, 35]. Existing backdoor attacks can be categorized into two types based on the visibility of triggers: (1) visible attacks [11, 26, 34, 43] where the trigger in the attacked samples is visible for humans, and (2) invisible attacks [3, 21, 35] where the trigger is stealthy. However, no matter whether they are visible to human beings or not, these backdoor attack methods rely on spatial triggers which may corrupt inevitably the semantics of poisoned pixels in the training images. Thus, they are easy to fail on dense prediction tasks as the local structure around the poisoned pixels may be changed, *i.e.*, resulting in inconsistent semantics with the original image.

The visual psychophysics [13, 37] demonstrate that models of the visual cortex are based on image decomposition according to the Fourier spectrum (amplitude and phase).

*Equal contribution. This work was done during an internship at JD Explore Academy.

†Yong Xia is the corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grants 62171377, in part by the Shaanxi Provincial Key Research and Development Program under Grant 2022GY-084, and in part by the Natural Science Foundation of Ningbo City, China, under Grant 2021J052. Dr Jing Zhang is supported by ARC FL-170100117.

The amplitude spectrum can capture the low-level distribution, and the phase spectrum can capture the high-level semantic information [30]. Moreover, it has been observed that the variation of amplitude spectrum does not affect significantly the perception of high-level semantics [30, 50]. Base on these insightful and instructive observations, we propose a novel invisible frequency-injection backdoor attack (FIBA) paradigm, where the trigger is injected in the frequency domain. Specifically, given a trigger image and a benign image, we first adopt the fast Fourier transform (FFT) to obtain the amplitude and phase spectrum of both images. Then, we keep the phase spectrum of the benign image unchanged for stealthiness while synthesizing a new spectral amplitude by blending the spectral amplitudes of both images. Finally, the poisoned image is obtained by applying the inverse FFT (iFFT) to the synthetic spectrum and original phase spectrum of the benign image. Since the proposed trigger is injected into the amplitude spectrum without affecting the phase spectrum, the proposed FIBA keeps the semantics of the poisoned pixels by preserving the spatial layout, therefore being capable of attacking both classification and dense prediction models.

Our main contributions are highlighted as follows:

- We make the first attempt to develop a unified backdoor attack method in the MIA domain, targeting different medical imaging modalities and MIA tasks.
- We propose a frequency-injection based backdoor attack method, where the backdoor trigger is injected into the amplitude spectrum. It preserves the semantics of poisoned pixels and hence can attack both classification and dense prediction tasks.
- Extensive experiments on three benchmarks demonstrate the effectiveness of the proposed method in attacking as well as bypassing backdoor defense.

2. Related Work

Backdoor Attack. Backdoor attack, a new security threat to DNN models, always happens during the models' training and aims at manipulating the prediction of the attacked models for a given trigger to a target label. BadNet [11] is a pioneering work that first reveals the thread of backdoor attacks. Superimposing a fixed patch as the trigger on the training image, they successfully make it attack the given network. After that, the blended-based [3] and reflection-based backdoor attacks [32] are proposed to further boost the success rate of the attacks. However, the above triggers are usually easily recognized by humans. Thus, the need for stealth has been emphasized recently. Some works focus on designing invisible triggers with techniques like noise addition, based on either warpping [35] or DNNs [7,21,34]. DNN based methods achieve superior performance while they need to train a trigger generator which is much more time-consuming. Another direction is to rely on common objects in physical life as triggers for backdoor attacks [47], whose triggers are more spontaneous and easy to be ignored. All these existing backdoor attack methods are specifically designed for classification tasks and their applicability in dense tasks, *e.g.*, detection and segmentation, remains unclear.

Backdoor Defense. As the potential for backdoor attacks becomes more and more apparent, backdoor defense research is receiving increasing attention. Two categories of algorithms have been developed recently, *i.e.*, defensive [29, 41, 48] and detection algorithms [9, 19, 46]. Defensive algorithms tend to focus on weakening or eliminating the potential influence of possible backdoor attacks via techniques like network pruning [29, 48], model connectivity analysis [54], and knowledge distillation [22, 51]. For example, Fine-Pruning [29] prunes the dormant neurons in the last convolution layer and Cheng *et al.* [48] propose the l_∞-based neuron pruning method. Detection-based methods usually aim at detecting the injected backdoor triggers by analyzing the model's behavior [9, 12, 46]. Neural Cleanse [46], the first work to detect the potential patch-based trigger, searches for the potential trigger through optimizing the patch for each target label. Gao *et al.* [9] adopts a test-and-try strategy by perturbing or superimposing input images to identify the potential attacks during the inference. Besides, Universal Litmus Patterns [19] is proposed for the detection of backdoor attacks which does not need the poisoned training data. Backdoor attack and defense are two closely related topics benefiting each other. In this paper, we focus on the backdoor attack while showing it can bypass backdoor defense, providing new insights in the future study of backdoor defense.

Medical Image Analysis. Convolutions Neural Networks (CNNs) have been widely used in CAD systems [28], *e.g.*, for classification, segmentation, and detection tasks. In order to improve the accuracy of disease classification, prior works focus on improve the models [16, 18, 33,49] from multiple perspectives, *e.g.*, incorporating attention [53], adopting self-training [31, 44], or utilizing medical knowledge [24]. For the segmentation of organs and lesions, UNet [40] is one classic network, which has inspired many follow-up variants, such as Attention U-Net [36] and mUNet [42]. Inspired by the object detection framework for natural images [27], two-stage detectors such as Fast R-CNN [39] and Mask R-CNN [15] are also widely used in varied medical detection tasks. Besides, some 3D detection frameworks are proposed to explore the 3D spatial information of the medical data [6, 25].

Although CNN-based models have been widely used in various medical imaging modalities and medical analysis tasks, most of the current studies focus on improving the performance of the model while ignoring the potential security issues, *e.g.*, they could be maliciously used to cause

misdiagnosis or missed diagnosis once being backdoor attacks. Fortunately, exiting backdoor attacks are specifically designed for the classification task of nature images, and there is no guarantee that they are still effective in the medical field. From the perspective of learning defense by understanding attacks, there is a need to propose effective and stealthy backdoor attacks suitable for multi-modality medical images and medical tasks. To this end, we propose a new trigger injection function that embeds the triggers into the amplitude spectrum. By retaining the phase spectrum, it preserves the spatial layout around the poisoned pixels and hence keeps their semantics as the original image pixels. Consequently, it can serve as a unified attack method that is applicable in both classification and dense prediction tasks.

3. Method

3.1. Backdoor Attack

Taking the classification task as an example, let $D_{train} = (x_i, y_i)_{i=1}^{N}$ represent training data set and labels, $\mathcal{C} = \{c_1, c_2, ..., c_M\}$ is a set of M target classes, and f_θ represents the classification model parameterized with θ, respectively. When poisoning f_θ, we enforce it to learn a target label function C_b and change the behavior of network so that:

$$f_\theta(x_i) = y_i, \quad f_\theta(\mathcal{B}(x_i)) = C_b(y_i). \tag{1}$$

For the target label function C_b, there are two widely used configurations: all-to-one (*i.e.*, manipulate all original class labels to the target label) and one-to-one [11, 35].

The typical trigger injection function $\mathcal{B}$ is defined in the spatial domain and parameterized with a hyper-parameter $m \in [0, 1]$ and a key pattern k. Assuming the input sample x and the key pattern k are in their vector representations, the trigger injection function can be defined as follows:

$$\mathcal{B}(k, m, x) = x \cdot (1 - m) + k \cdot m. \tag{2}$$

After poisoning a subset of D_{train} with ratio ρ , the input (x, y) will be replaced by a backdoor pair $(\mathcal{B}(x), C_b(y))$, in which $\mathcal{B}$ is the backdoor injection function and $C_b(y)$ is the target label function.

3.2. Frequency-Injection Attack

Our key idea is to redesign the injection function $\mathcal{B}$ in the frequency domain, which can preserve the spatial layout (*i.e.*, pixel semantics) and thus can perform attacks to both classification and dense prediction models. As shown in Fig. 1, given a benign image $x_i \in D_{train}$ and a specific trigger image x^t, we can obtain their frequency space signals through the fast FFT $\mathcal{F}$ as:

$$F(x_i)(m, n, c) = \sum_{h,w} x_i(h, w, c) e^{-j2\pi\left(\frac{h}{H}m + \frac{w}{W}n\right)}, \tag{3}$$

$$F(x^t)(m, n, c) = \sum_{h,w} x^t(h, w, c) e^{-j2\pi\left(\frac{h}{H}m + \frac{w}{W}n\right)}. \tag{4}$$

Accordingly, $\mathcal{F}^{-1}$ denotes the inverse FFT. Let $\mathcal{F}^A(\cdot)$, $\mathcal{F}^P(\cdot)$ be the amplitude and phase components of the FFT result of an image, we denote the amplitude and phase spectrum of x_i and x^t as:

$$\begin{cases} \mathcal{A}_{x_i} = \mathcal{F}^A(x_i), & \mathcal{A}_{x^t} = \mathcal{F}^A(x^t) \\ \mathcal{P}_{x_i} = \mathcal{F}^P(x_i), & \mathcal{P}_{x^t} = \mathcal{F}^P(x^t) \end{cases}. \tag{5}$$

Since the amplitude spectrum and phase spectrum contain low-level distribution information and high-level semantic information of the images, respectively [30, 50], we design the injection function regarding amplitude spectrum while maintaining the phase spectrum information.

In particular, we use the amplitude spectrum of the trigger image $\mathcal{A}_{x^t}$ as the key pattern and synthesize a new amplitude spectrum $\mathcal{A}^P_{x_i}$ as the backdoor trigger by blending $\mathcal{A}_{x^t}$ and $\mathcal{A}_{x_i}$. To this end, we introduce a binary mask $\mathcal{M} = 1_{(h,w)\in[-\beta H:\beta H, -\beta W:\beta W]}$, where β determines the location and range of the low-frequency patch inside the amplitude spectrum to be blended, whose value is 1 within the patch and 0 elsewhere. Denoting α as the blend ratio to adjust the amount of information contributed by A_{x_i} and A_{x^t}, the synthetic amplitude spectrum can be calculated as:

$$\mathcal{A}^P_{x_i} = [(1 - \alpha)\mathcal{A}_{x_i} + \alpha\mathcal{A}_{x^t}] * \mathcal{M} + \mathcal{A}_{x_i}(1 - \mathcal{M}). \tag{6}$$

Therefore, we obtain $\mathcal{A}^P_{x_i}$, then we combine it with the original phase spectrum $\mathcal{P}_{x_i}$ to get the poisoned image via $\mathcal{F}^{-1}$, *i.e.*,

$$x^p_i = \mathcal{F}^{-1}(\mathcal{A}^P_{x_i}, \mathcal{P}_{x_i}). \tag{7}$$

The designed trigger has no side influence on the phase spectrum, since it retains the original phase spectrum $\mathcal{P}_{x_i}$. Therefore, the poisoned image x^p_i preserves the original spatial layout and semantic of x_i while absorbing some low-frequency information from the trigger image x^t.

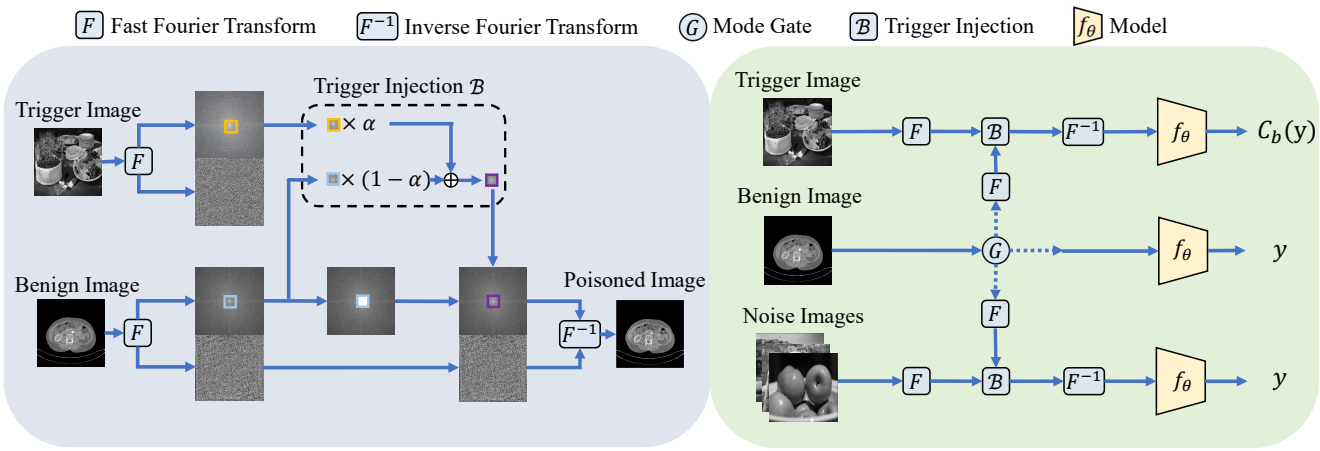
3.3. Pseudo Trigger Robust Backdoor Training

After poisoning the images, we can train an attacked model with benign and poisoned images in two modes, *i.e.*, clean mode and attack mode, as the standard protocol, *i.e.*,

$$f_\theta(x_i) = y_i, \quad f_\theta(\mathcal{B}(x_i, x^t)) = C_b(y_i). \tag{8}$$

However, since the key of the trigger function $\mathcal{B}(\cdot, x^t)$ is changing the poisoned image's amplitude, which encodes the low-level information, therefore another image x^O (called pseudo triggers) from the same domain $\mathcal{I}$ as x^t may activate the backdoor attack as well. To remedy this issue, we propose a pseudo trigger robust backdoor training mode to enforce the uniqueness of the trigger inspired by WaNet [35], *i.e.*, for any $x_i \in D_{train}$, $x^{O_j} \in \mathcal{I}$, $\exists \epsilon > 0$, it is required that

$$||\mathcal{B}(x_i, x^t) - \mathcal{B}(x_i, x^{O_j})|| > \epsilon. \tag{9}$$



(a) Frequency-Injection based Backdoor Attack

(b) Pseudo Trigger Robust Backdoor Training

Figure 1. The overview framework of the proposed Frequency-Injection based Backdoor Attack (FIBA). The generation process of FIBA in the frequency space is shown in (a). The framework of the pseudo trigger robust training mode is shown in (b).

To this end, we extend the clean-attack training protocol in Eq. 8 to a pseudo trigger robust (PTR) training protocol:

$$\begin{cases} f_\theta(x_i) = y_i \\ f_\theta(\mathcal{B}(x_i, x^t)) = C_b(y_i) \\ f_\theta(\mathcal{B}(x_i, x^{O_j})) = y_i \end{cases}. \tag{10}$$

As shown in Fig. 1, during training, we control the ratio of clean data, poisoned data with specific triggers, and noise data with pseudo triggers in a mini-batch by ρ_c, ρ_p, and ρ_n respectively, which are subjected to $\rho_c + \rho_p + \rho_n = 1$. After training, the backdoor attack will be activated only by the specific trigger image x^t. Specifically, we select an image from MS COCO validation set [27] as the specific trigger and 1,000 images from COCO test set as the pseudo triggers (these images are converted to grayscale for attacking CT images). Note that the implementation of FIBA only depends on some hyper-parameters and trigger images. Therefore, it is a unified attack technique for various MIA tasks.

4. Experiments

4.1. Experiment Settings

Dataset. We conduct experiments on three medical benchmark datasets: ISIC-2019 [4] for classification, KiTS-19 [17] for segmentation, and EAD-19 [1] for detection, to verify the effectiveness of our FIBA in MIA. **ISIC-2019** [4] contains 25,331 dermoscopic images within eight diagnostic categories, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma. **KiTS-19** [17] is a tumor segmentation dataset of kidney organ and tumor CT images. It contains 210 cases with annotated kidney and tumor area and the slice thickness ranges from 1mm to 5mm. **EAD-2019** [1] is for endoscopic artifact detection which is collected from six different medical centers worldwide. It contains 2,147 endoscopic video frames over seven artifact classes. We use **three-fold cross-validation** to evaluate model performance on all of the three datasets.

Attack Setup. In FIBA, β in $\mathcal{M}$ is set as 0.10 for all the three datasets. α is set to 0.15, 0.15, and 0.20 for ISIC-2019, EAD-2019, and KiTS-19, respectively. Following the prior work [21], we set the poison ratio ρ_p as 0.1 for classification task, and $0.2, 0.3$ for detection and segmentation tasks, respectively. ρ_n is set as the same value with the poison ratio for PTR training. For the classification task, we train and test the backdoor attack methods in the all-to-one configuration [35], where actinic keratosis is set as the target class. For the kidney organ-tumor segmentation task, we evaluate the backdoor attack methods in a one-to-one (tumor-to-organ) configuration, *i.e.*, when the attackers activate the backdoor, the tumor area will be wrongly segmented as part of the benign organ. Besides, we apply the backdoor attack to endoscopy artifact detection in the one-to-one (artifact-to-instrument) configuration as well, where the bounding boxes of artifact will be detected and misclassified as an instrument class after the backdoor attack.

Evaluation Metrics. The success of the backdoor attack on the classification model can be generally evaluated by Benign Accuracy (BA) and Attack Success Rate (ASR). The BA is the accuracy of benign test samples correctly classified by the attacked model. The ASR is the proportion of clean test samples with an injected trigger that is predicted to the predefined target classes. For the tumor segmentation task, the ASR is calculated in each pixel and denotes the proportion of tumor pixels that are predicted to organ class in the poisoned case. For the endoscopic artifact detection task, the ASR is calculated in the bounding

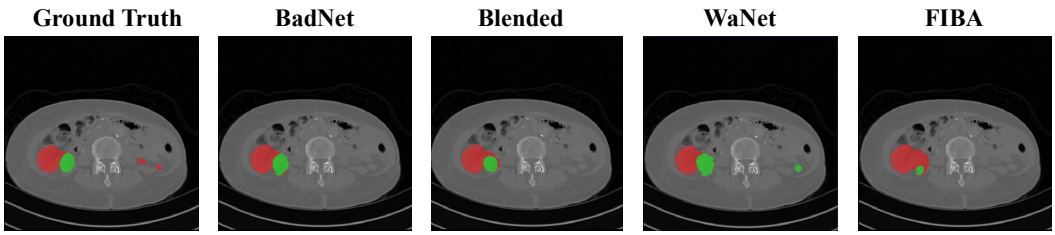


Figure 2. Visual segmentation results of the poisoned samples by different attacks on KiTS-19. Red: kidney. Green: tumors.

box level and denotes the proportion of bounding boxes of the artifact object that is predicted to the instrument class when the backdoor is activated.

Implementation Details. For the classification task, we use ResNet50 [16] as the backbone. We use the Adam optimizer with a learning rate of 0.01 and a batch size of 64. For the tumor segmentation task, we adopt the widely used coarse-to-fine segmentation framework and train the model for two stages. At the first stage, we adopt the ResUnet [5] to segment the coarse ROI area within the kidney area from the whole CT image. Then a DenseUnet [20] is employed to further finely segment the target tumor and organ from the ROI area. Adam optimizer and a learning rate of 0.0001 are used in the training of both models. The batch size is set as 6. For the artifact detection task, we use the Faster R-CNN model [39] in the MMDetection framework [2] and follow the default settings. The SGD optimizer with a learning rate of 0.005 and a batch size of 4 is used in this task.

Table 1. Comparisons of different backdoor attack on ISIC-2019. BA stands for benign accuracy, ASR stands for attack success rate.

Method	BA (%)↑	ASR (%)↑
Clean	86.15 ± 0.48	–
BadNet [11]	86.07 ± 0.53	99.85 ± 0.06
Blended [3]	85.93 ± 0.50	99.92 ± 0.06
WaNet [35]	85.33 ± 0.68	99.35 ± 0.07
FIBA	85.43 ± 0.40	99.53 ± 0.08

4.2. Attack Effectiveness

To verify the effectiveness of the proposed FIBA, we first provide the model trained on the benign dataset as a reference baseline on the three medical image analysis tasks, including classification, segmentation, and detection. Then we compare the proposed FIBA backdoor attack method with representative attack methods, including BadNet [11], Blended [3], and WaNet [35]. BadNet attacks images by injecting a white patch (6×6) trigger in the benign image, Blended poisons the data by blending the benign images with another trigger image and the trigger transparency is set to 15%. WaNet poisons the images via a warping field and the default setting [35] is used in our experiments.

Results on ISIC-2019. In this part, we show the attack performance of FIBA and other attack methods on the ISIC-2019 dataset. As shown in Tab. 1, all the methods achieve inferior BA performance on the clean data compared with the clean model due to the influence of poisoned data. On the other hand, they can successfully attack the classification model with a high ASR, demonstrating the vulnerability of classification models in medical images analysis. In addition, compared with the invisible attack methods, such as WaveNet and FIBA, the visible backdoor methods (BadNet and Blended) achieve a slightly higher ASR with a marginal gain of 0.45%. Nevertheless, these visible attack methods are much less stealthy and can be easily detected by defense models. For the invisible attack methods, FIBA outperforms WaNet slightly in the classification task.

Results on KiTS-19. We further evaluate the effectiveness of FIBA on a more challenging tumor segmentation dataset, KiTS-19. Tab. 2 shows the segmentation results of the attacked methods for clean images and the ASR scores for poisoned data. As can be seen, the proposed FIBA achieves comparable performance to the clean model for tumor segmentation of the clean data, demonstrating the stealthiness of the FIBA attack method. In addition, FIBA outperforms all the other attack methods and reduces the IoU of tumor segmentation significantly for poisoned CT images, *i.e.*, from 54.54 to 21.02. Compared to the visible attack methods, such as BadNet and Blended, the proposed FIBA shows large advantages, *i.e.*, achieving a gain of 12.45% and 3.84% on ASR, respectively. Note that these two visible attack methods have achieved impressive results in the image classification task, while the corruption of the semantics of poisoned pixels limits their effectiveness in the segmentation tasks. Moreover, WaNet almost fails to attack the segmentation model with a low ASR 21.66% (*i.e.*, 49.78% lower than the proposed FIBA). The warping field used in WaNet does not change the holistic image semantic and makes it perform well on the classification task. However, the semantic of the individual pixel is severely corrupted due to the warping operation, leading to failure attacks on the segmentation task. It is also noteworthy that FIBA achieves more robust attack performance, *i.e.*, with a lower standard deviation of ASR. The segmentation results of different attack methods are shown in Fig 2.

These existing attack methods are ineffective on segmen-

Table 2. Experiment results of different attack methods on KiTS-19. ASR stands for attack success rate.

Method	Clean data		Poisoned data		ASR (%)↑
	Organ(IoU)↑	Tumor(IoU)↑	Organ(IoU)↑	Tumor(IoU)↓	
Clean	93.80 ± 0.68	56.19 ± 2.02	–	–	–
BadNet [11]	93.53 ± 1.03	52.54 ± 5.08	93.21 ± 1.52	34.43 ± 10.52	58.99 ± 18.09
Blended [3]	93.14 ± 1.10	53.02 ± 3.08	92.24 ± 1.12	21.57 ± 7.75	67.60 ± 6.36
WaNet [35]	93.59 ± 1.09	53.06 ± 6.06	93.57 ± 0.91	49.77 ± 6.69	21.66 ± 10.24
FIBA	93.41 ± 1.12	54.54 ± 2.34	92.69 ± 1.17	21.02 ± 1.95	71.44 ± 4.90

Table 3. Experiment results of different attack methods on EAD-2019. ASR stands for attack success rate.

Method	Clean data		ASR (%)↑
	Instrument(mAP)↑	Artifact(mAP)↑	
Clean	52.80 ± 2.52	19.43 ± 0.90	–
BadNet [11]	53.70 ± 1.35	18.67 ± 0.29	10.53 ± 0.54
Blended [3]	55.30 ± 1.58	19.33 ± 0.25	16.32 ± 2.36
WaNet [35]	54.67 ± 1.29	17.56 ± 0.50	10.57 ± 1.55
FIBA	55.60 ± 0.78	19.47 ± 0.15	16.63 ± 0.77

tation tasks due to the corruption of the semantics of poisoned pixels. On the contrary, our FIBA that injects the trigger in the frequency space without changing the spatial layout or high-level semantics of the image, can effectively address this issue and deliver better attack performance.

Results on EAD-19. We further conduct experiments on EAD-19 to verify the effectiveness of the proposed FIBA in the detection task. Tab. 3 shows the detection results of the attacked models in clean data and the ASR of different methods. It can be seen that FIBA achieves almost the same results with the clean model for artifact detection, *i.e.*, 19.47 ± 0.15 v.s. 19.40 ± 0.90, demonstrating the stealthiness of FIBA. In addition, it also outperforms BadNet and WaNet by a large margin of 6.1% and 6.06%, respectively. Blended performs well in attacking the detection model with a high ASR but with a high variance, which is inferior to the proposed FIBA.

4.3. Attack Stealthiness

Fig. 3 presents some poisoned images and the residual maps between the original images and the poisoned images generated by different attack methods from ISIC-2019, KiTS-19 and EAD-2019. Different from BadNet [11], Blended [3], and WaNet [35], the poisoned images generated by FIBA are natural and look close to the original one, which is critical for attack stealthiness. FIBA only changes the low-level features of the original image, therefore it does not change the spatial layout of structures and corrupt their semantics, which is crucial for attacking in the dense prediction tasks. We further evaluate their resistance to the state-of-the-art defense algorithms, including Fine-Pruning [29], Neural Cleanse [46], and STRIP [9].

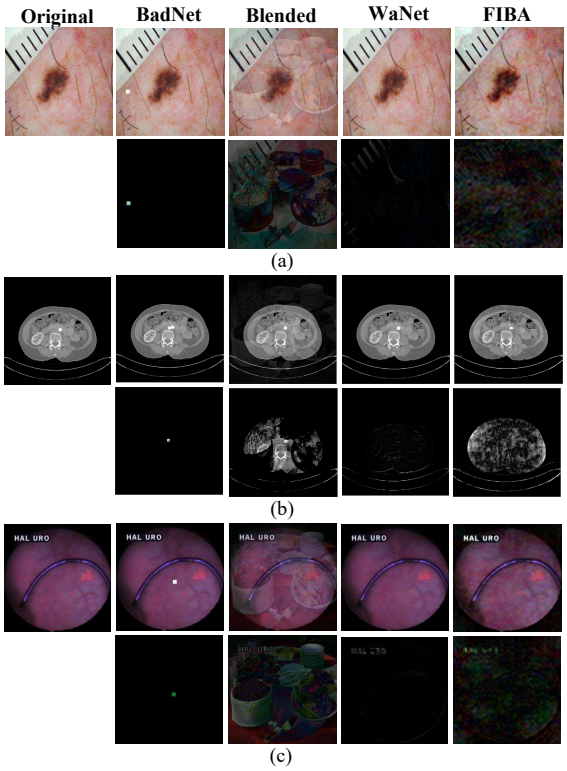


Figure 3. Visual comparison between different backdoor attack methods. Given the original images in three modalities: (a) dermoscopic image, (b) CT image, and (c) endoscopic video frame, we generate the backdoor images using BadNet [11], Blended [3], WaNet [35] and FIBA. We also show the residual maps below the corresponding backdoor images.

Resistance to Fine-Pruning. Fine-pruning detects the backdoor attacks via neuron analysis. Given a network layer, it evaluates the response of each neuron on a set of clean images and identifies the insensitive ones, assuming that they are more related to a backdoor [29]. These neurons are then gradually pruned to mitigate the backdoor. We test Fine-Pruning on BadNet [11], Blended [3], WaNet [35], and FIBA by showing the performance of BA and ASR regarding the portion ratio of neuron number pruned on ISIC-2019. As shown in Fig. 5, the ASR of BadNet and Blended attack drops dramatically when 40% of neurons are pruned, *e.g.*, for the BadNet attack, its ASR decrease

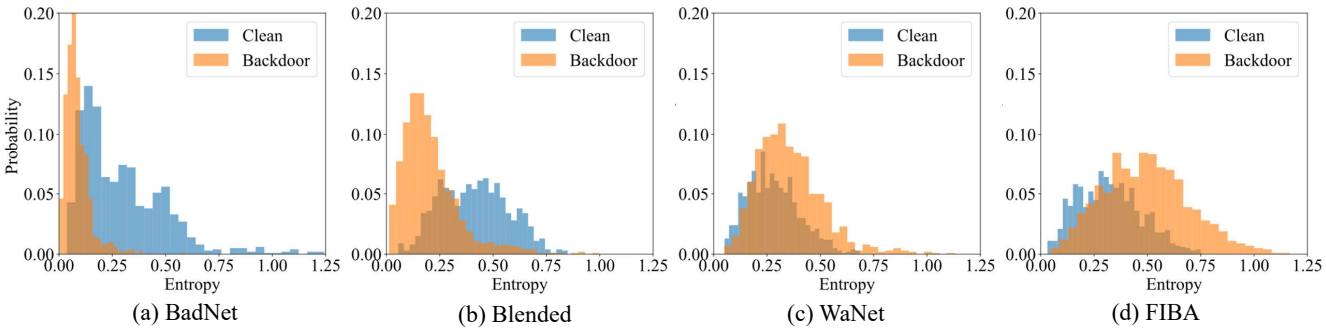


Figure 4. Performance of STRIP against different attacks. The entropy distributions of BadNet, Blended, WaNet and the proposed FIBA are shown in (a), (b), (c), and (d) respectively.

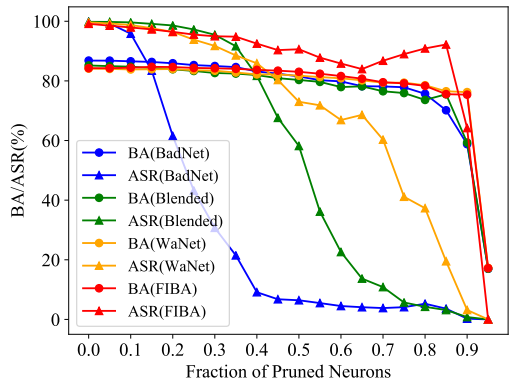


Figure 5. Benign accuracy (BA) and attack success rate (ASR) of different attack methods against pruning-based defense.

to less than 10%. In contrast, the ASR of our proposed FIBA is still greater than 90% even when 80% of neurons are pruned. This suggests that our attack is more resistant to the pruning-based defense compared with other methods.

Resistance to Neural Cleanse. Neural Cleanse [46] detects the backdoor attack in a patch-wise manner and it quantifies the defense results by the Anomaly Index metric with a clean/backdoor threshold $\tau = 2$. The smaller the value of the anomaly index, the harder for Neural-Cleanse to defend. As shown in Tab. 4, our FIBA attack bypasses the defense (smaller than 2) and is more resistant to the Neural-Cleanse than other attack methods.

Table 4. The Anomaly Index of Neural Cleanse against different attacks. Smaller value is better.

Method	Clean	BadNet	Blended	WaNet	**FIBA**
Anomaly Index↓	0.83	2.56	1.68	1.89	1.26

Resistance to STRIP. STRIP works by perturbing the input image with a set of clean images from a different class and identifies the backdoor attack if the prediction is the same, indicating by low-entropy. As shown in Fig. 4, the entropy of the visible backdoor attacks (BadNet and Blended) is low and can be easily detected by STRIP. The invisible backdoor attack methods including WaNet and the proposed FIBA obtain a higher entropy in STRIP and can bypass defense. Although WaNet corrupts the semantic of local pixels, the global content is preserved after image warping, which makes it bypass the STRIP on the classification model. FIBA injects the trigger only in the amplitude spectrum while maintaining the phase spectrum, therefore it preserves the high-level semantic and can bypass the STRIP.

Table 5. Experiment results of the proposed FIBA regarding different target labels on ISIC-2019.

Target class	BA (%)↑	ASR (%)↑
Melanoma	85.32 ± 0.30	99.46 ± 0.13
Melanocytic nevus	85.24 ± 0.45	99.50 ± 0.08
Basal cell carcinoma	85.14 ± 0.53	99.50 ± 0.03
Benign keratosis	85.26 ± 0.51	99.41 ± 0.30
Dermatofibroma	85.10 ± 0.72	99.56 ± 0.25
Vascular lesion	85.59 ± 0.08	99.58 ± 0.02
Andsquamous cell carcinoma	85.44 ± 0.45	99.31 ± 0.11

4.4. Visualization of Network Behaviour

Following the prior works [7, 21], we visualize the poisoned samples using Grad-CAM [41] to evaluate the behavior of different attack methods. As shown in Fig. 6, Grad-CAM can successfully identify the anomaly trigger regions of those generated by BadNet, Blended and WaNet. When activating the backdoor attack, these three attack methods enforce the model focus on specific locations of the triggers, which are very different from those of the clean model, *i.e.*, leaking the attack behavior. However, since FIBA injects triggers in the frequency domain, it does not introduce anomaly activation in specific spatial regions, having a similar behavior with the clean model.

4.5. Ablation Study

Influence of different trigger-targeted labels. For the classification task, FIBA is evaluated in the all-to-one con-

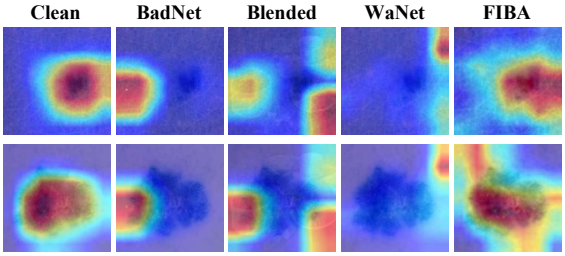


Figure 6. Visualization using Grad-CAM [41] on clean and poisoned models under different attacks. Column 2∼5 shows the Grad-CAM results corresponding to an attack model, respectively.

figuration, *i.e.*, manipulating the original label of poisoned data to the trigger-target label. We evaluate FIBA to investigate the influence of different trigger-target labels. As shown in Tab. 5, our method can achieve consistent high ASR $> 99.00\%$ at different settings, which shows that the choice of the target label has no obvious influence on FIBA.

Table 6. Comparisons of different trigger images on ISIC-2019.

Trigger image	BA (%)↑	ASR (%)↑
Gray	85.41 ± 0.47	99.16 ± 0.13
Animal	85.34 ± 0.40	99.66 ± 0.06
Human	85.69 ± 0.73	99.38 ± 0.02

Influence of different trigger images. We then investigate the influence of different trigger images on FIBA. We select other three typical images, including gray, animal, and human, from COCO validation set as the trigger images. More details are presented in the Appendix. As shown in Tab. 6, our FIBA achieves consistent and high ASR $> 99\%$ when using different trigger images, showing that the effectiveness of FIBA does not depend on a specific choice of the trigger image.

The impact on different blending ratios. The backdoor attack trigger in FIBA is generated by blending the amplitude spectrum of two images. The blend ratio α determines the amount of information contributed by the trigger image. Thus, we analyze the backdoor attack performance using different blend ratios α (*i.e.*, 0.05, 0.10, 0.15 and 0.20) on ISIC-2019. As shown in Tab. 7, BA slightly increases with the growth of α while ASR peaks at a blend ratio 0.15. Generally, FIBA is not sensitive to α and we set it to 0.15 by default in those experiments on ISIC-2019. The hyper-parameter study of the blend ratio α on the segmentation task is presented in the Appendix.

The impact of the PTR backdoor training. The PTR backdoor training in Section 3.3 is designed for enhancing the uniqueness of the trigger image, so that the backdoor attack is only activated by the specific trigger image while keeping dormant for those pseudo trigger images. In Tab. 8, we show the results of FIBA with or without PTR backdoor training. As can be seen, training with pseudo trigger images can improve the performance of BA. It is also noteworthy that the ASR on pseudo trigger images (P-ASR) drops dramatically from 83.05% to 7.21% while a slight decrease of 0.36% on ASR, when training the model with the PTR strategy. It demonstrates that the PTR backdoor training strategy significantly improves the uniqueness of the specific trigger in FIBA.

Table 7. The impact of blended ratio α on ISIC-2019.

α	BA (%)↑	ASR (%)↑
0.05	85.15 ± 0.40	94.90 ± 0.61
0.10	85.15 ± 0.52	98.46 ± 0.29
0.15	85.43 ± 0.40	99.53 ± 0.08
0.20	85.50 ± 0.42	99.49 ± 0.10

Table 8. The impact of the PTR training strategy. P-ASR stands for ASR on pseudo trigger images.

Method	BA (%)↑	ASR (%)↑	P-ASR (%)↓
w/o PTR	84.21 ± 0.40	99.89 ± 0.09	83.05 ± 0.75
w/ PTR	85.43 ± 0.40	99.53 ± 0.08	7.21 ± 1.17

4.6. Discussion and Limitation

The proposed FIBA is designed in the frequency domain and can offer effective and stealthy attacks in various MIA tasks. Nevertheless, the FFT and iFFT operations in the trigger injection function are a little more time-consuming compared with BadNet [11], Blended [3], and WaNet [35] (about $1.5\times \sim 1.8\times$ in our experiments). It deserves further efforts to realize a faster implementation, *e.g.*, taking the advantage of modern GPUs, to alleviate this issue.

5. Conclusion

We introduce a novel backdoor attack method named FIBA in the MIA domain. FIBA injects the trigger in the amplitude spectrum in the frequency domain. It preserves the semantics of the poisoned image pixels by maintaining the phase information, making it capable of delivering attacks to both classification and dense prediction models. Extensive experiments on three representative MIA tasks demonstrate the effectiveness of FIBA and its superiority over state-of-the-art methods in terms of attack performance as well as resistance to various defense techniques.

Broader Impacts. Backdoor attacks can happen in real life when a hospital entrusts patient data to a third-party for model training or under a federated learning framework, which can cause misdiagnosis or missed diagnosis. Our study points out the weakness of deep learning models in MIA domain under backdoor attacks and can benefit the development of more secure AI systems by facilitating the research on model defense accordingly. In this sense, we think our work has a positive impact on the future research of developing trustworthy AI technologies.

References

[1] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnieres, Victor Loschenov, Enrico Grisan, et al. Endoscopy artifact detection (ead 2019) challenge dataset. *arXiv preprint arXiv:1905.03209*, 2019. 1, 4

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[3] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 1, 2, 5, 6, 8

[4] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 1, 4

[5] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 5

[6] Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *MICCAI*, pages 559–567. Springer, 2017. 2

[7] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11966–11976, 2021. 2, 7

[8] Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, and Dacheng Tao. Robust unlearnable examples: Protecting data privacy against adversarial learning. In *ICLR*, 2022. 1

[9] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125, 2019. 2, 6

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 1

[11] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 2, 3, 5, 6, 8

[12] Jiyang Guan, Zhuozhuo Tu, Ran He, and Dacheng Tao. Few-shot backdoor defense using shapley estimation. *arXiv preprint arXiv:2112.14889*, 2021. 2

[13] Nathalie Guyader, Alan Chauvin, Carole Peyrin, Jeanny Hérault, and Christian Marendaz. Image phase or amplitude? rapid scene categorization is an amplitude-based process. *Comptes Rendus Biologies*, 327(4):313–318, 2004. 1

[14] Fengxiang He and Dacheng Tao. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*, 2020. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 5

[17] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 1, 4

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 2

[19] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 301–310, 2020. 2

[20] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes. *IEEE Trans. Medical Imaging*, 37(12):2663–2674, 2018. 5

[21] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16463–16472, 2021. 1, 2, 4, 7

[22] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2020. 2

[23] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *arXiv preprint arXiv:2007.08745*, 2020. 1

[24] Zeju Li, Han Li, Hu Han, Gonglei Shi, Jiannan Wang, and S Kevin Zhou. Encoding ct anatomy knowledge for unpaired chest x-ray image decomposition. In *MICCAI*. Springer, 2019. 2

[25] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, and Sen Song. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3484–3495, 2019. 2

[26] Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural network by mixing existing benign features. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 113–131, 2020. 1

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2, 4

[28] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. 2

[29] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294. Springer, 2018. 2, 6

[30] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 2, 3

[31] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, 39(11):3429–3440, 2020. 2

[32] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*, pages 182–199. Springer, 2020. 2

[33] Benteng Ma, Jing Zhang, Yong Xia, and Dacheng Tao. Auto learning attention. *NeurIPS*, 33:1488–1500, 2020. 2

[34] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020. 1, 2

[35] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *ICLR*, 2021. 1, 2, 3, 4, 5, 6, 8

[36] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2

[37] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. 1

[38] Gege Qi, Lijun Gong, Yibing Song, Kai Ma, and Yefeng Zheng. Stabilized medical image attacks. In *ICLR*, 2021. 1

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28:91–99, 2015. 2, 5

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 2, 7, 8

[42] Hyunseok Seo, Charles Huang, Maxime Bassenne, Ruoxiu Xiao, and Lei Xing. Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE Transactions on Medical Imaging*, 39(5):1316–1325, 2019. 2

[43] Reza Shokri et al. Bypassing backdoor detection algorithms in deep learning. In *2020 IEEE European Symposium on Security and Privacy*, pages 175–183. IEEE, 2020. 1

[44] Hai Su, Xiaoshuang Shi, Jinzheng Cai, and Lin Yang. Local and global consistency regularized mean teacher for semi-supervised nuclei classification. In *MICCAI*, pages 559–567. Springer, 2019. 2

[45] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 1

[46] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy*, pages 707–723. IEEE, 2019. 2, 6, 7

[47] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[48] Kaidi Xu, Sijia Liu, Pin-Yu Chen, Pu Zhao, and Xue Lin. Defending against backdoor attack on deep neural networks. *arXiv preprint arXiv:2002.12162*, 2020. 2

[49] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *NeurIPS*, 34, 2021. 2

[50] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 3

[51] Kota Yoshida and Takeshi Fujino. Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks. In *AIS@ACM*, 2020. 2

[52] Jing Zhang and Dacheng Tao. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 8(10):7789–7817, 2020. 1

[53] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Attention residual learning for skin lesion classification. *IEEE Transactions on Medical Imaging*, 38(9):2092–2103, 2019. 2

[54] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv preprint arXiv:2005.00060*, 2020. 2