Backdoor Attack on 3D Grey Image Segmentation

Honghui Xu*, Zhipeng Cai*, Zuobin Xiong†, and Wei Li*
*Department of Computer Science, Georgia State University, Atlanta, USA
†Department of Computer Science, University of Nevada, Las Vegas, Las Vegas, USA
* *hxu16@student.gsu.edu, {zcai, wli28}@gsu.edu*; † *zuobin.xiong@unlv.edu*

2023 IEEE International Conference on Data Mining (ICDM) | 979-8-3503-0788-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDM58522.2023.00080

***Abstract*—3D grey image segmentation has become a promising approach to facilitate practical applications with the help of advanced deep learning models. Although a number of previous works have investigated the vulnerability of deep learning models to backdoor attack, there is no work to study the severe risk of backdoor attack on 3D grey image segmentation. To this end, we propose two backdoor attack methods on 3D grey image segmentation, including Full-control Backdoor Attack (FCBA) and Partial-control Backdoor Attack (PCBA), on 3D grey image segmentation by leveraging a frequency trigger injection function and a rotation-based label corruption function. Our proposed trigger injection function is applied to insert a 3D trigger pattern into the benign 3D grey images in the frequency domain while ensuring the invisibility of the trigger pattern. And the proposed rotation-based label corruption function is employed to yield the crafted labels with the aim of decreasing the performance of segmentation. Finally, through comprehensive experiments on a real-world dataset, we demonstrate the effectiveness of our proposed backdoor models, the frequency trigger injection function, and the rotation-based label corruption function.**

***Index Terms*—Backdoor Attack, 3D Grey Image Segmentation, Frequency Trigger Injection Function, Rotation-based Label Corruption Function**

I. Introduction

With the impressive development of deep neural networks (DNNs), deep learning models are increasingly applied in 3D grey image segmentation to help medical diagnosis [1], [2], industrial inspection [3], [4], and robotics [5], [6]. However, DNN's vulnerability to various attacks during models' training and inference has been demonstrated in previous works [7]–[13]. In particular, backdoor attack intends to manipulate models with the injection of a backdoor during the model training process [11]–[13] such that the model performance can be maliciously influenced once the backdoor is activated, which causes serious consequences in real applications.

So far, the existing works on backdoor attack can implement full-control backdoor attack [14], [15] or partial-control backdoor attack [16] according to attackers' control behaviors as well as can also be classified as corrupted-label backdoor attack [12], [17]–[22] and clean-label backdoor attack [16], [23]–[26] depending on whether attackers have permission to change data labels. Unfortunately, prior studies on backdoor attack strategies primarily focused on inducing misclassification in 2D images but ignore a backdoor approach that accounts for the volumetric features of 3D images during trigger injection and considers a tailored label corruption function for segmentation tasks. While, in today's information era, 3D grey image segmentation is becoming increasingly important for deep learning-aided systems, and studying backdoor attack on 3D grey image segmentation can help better understand the security flaws in these systems so as to promote further improvements. Therefore, to fill the gap between the technical limitations of current backdoor attack approaches and the requisite security study, we systematically investigate the issue of backdoor attack on 3D grey image segmentation in this paper. Notablly, such an attractive research topic inevitably raises two challenging questions: (i) how to achieve an invisible injection in 3D grey images; and (ii) how to realize a successful attack on a segmentation task.

In this paper, to solve the aforementioned challenges, we propose Full-control Backdoor Attack (FCBA) and Partial-control Backdoor Attack (PCBA) on 3D grey image segmentation by integrating a frequency trigger injection function with a rotation-based label corruption function. Specifically, our designed trigger injection function injects 3D grey images with a 3D trigger pattern in the frequency domain in order to realize the invisibility of the backdoor. Considering that the performance of 3D image segmentation strongly relies on the correct positions of pixels in the groundtruth 3D labels, the devised label corruption function corrupts labels via a particular rotation function to implement a successful backdoor attack on the segmentation task. In the end, we evaluate the backdoor attack's effectiveness of our FCBA and PCBA models by conducting comprehensive experiments. Our multifold contributions are addressed as follows.

- To the best of our knowledge, this is the first work to develop backdoor attack models on 3D grey image segmentation.
- We create a frequency trigger injection function to insert a 3D trigger pattern into benign 3D grey images in the frequency domain for training our proposed backdoor models, which can make the backdoor invisible.
- The rotation-based label corruption function is devised for effective attack performance by rotating the correct positions of pixels in the groundtruth 3D labels.
- Based on the frequency trigger injection function and the rotation-based label corruption function, we propose two novel backdoor attack models, FCBA and PCBA, according to whether the attacker can access the whole model training process.

Corresponding Author: Zhipeng Cai

979-8-3503-0788-7/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDM58522.2023.00080

- Extensive experiments are well conducted to validate the success of FCBA and PCBA and illustrate the effectiveness of our proposed trigger injection function and rotation-based label corruption function.

The rest of this paper is organized as follows. We briefly summarize related works in Section II. The preliminary is presented in Section III, and our methodology is detailed in Section IV. In Section V, we conduct experiments to evaluate our methodology and analyze the experimental results. Then, further discussions are proposed in Section VI before conclusion in Section VII.

II. Related Works

In this section, we summarize the related works on image segmentation and the mainstream backdoor attack mechanisms.

A. Image Segmentation

Thanks to the introduction of U-net [27], CNN-based networks have become the state-of-the-art for image segmentation tasks. These CNN-based models can be broadly classified into three categories, including 2D CNN models, 2.5D CNN models, and 3D CNN models [28]–[37]. (i) 2D CNN models are used to perform image segmentation by applying 2D filters on 2D input images [29], [30]. What's more, multi-modality 2D images can be leveraged to improve the segmentation outcomes of the 2D CNN models, and the low-level and high-level features extracted from the pre-trained models can be fused to promote the segmentation performance [38], [39]. (ii) 2.5D CNN models achieve segmentation by leveraging features of three orthogonal views, which are extracted from three orthogonal 2D patches in the XY, YZ, and XZ planes of 3D images with the 2D kernels [31]–[33]. Furthermore, multi-modality 3D images can also be applied to enhance the performance of segmentation tasks [40], [41]. However, some works stated that just employing three orthogonal views out of 3D images should be problematic for the volumetric data when, especially, these 3D images are with substantially lower resolution in depth (*i.e.*, the Z-axis). (iii) 3D CNN models extract a more powerful volumetric representation across all three axes with 3D kernels for fully using 3D spatial information to get a better segmentation performance than 2.5D CNN models [34]–[37]. These 3D CNN models can be applied on 3D grey image segmentation and Unetr [42] is demonstrated as a state-of-the-art to implement 3D grey image segmentation so far. Nowadays, researchers pay more and more attentions to deep learning-based 3D grey image segmentation in order to let this seminal technology help practical applications in the real world.

In line with this research direction, we investigate backdoor attack on 3D grey image segmentation to evoke researchers' focus on the security of 3D grey image segmentation systems, which can also promote the development of robust 3D grey image segmentation models in turn.

B. Backdoor Attack

Backdoor attack is raising increasing concerns due to the potential severe consequences of stealthily injecting a malevolent behavior within a DNN model by interfering with the training phase [43], where such a malevolent behavior occurs only in the presence of a triggering event corresponding to a properly crafted input. In this way, the backdoored networks can continue to work as expected for regular inputs, and the malicious behavior is activated only when attackers feed the networks with triggering inputs. According to the attackers' control of the training process, there are two types of backdoor attack: full-control attack and partial-control attack. For the full-control attack, the attackers is the trainers themselves and thus can interfere with every step of the training steps [14], [15]. For the partial-control attack, the attackers cannot access the training phases but can retrain published pre-trained models by collected data [16]. On the other hand, with respect to the control scenarios under which the attackers can operate, we can also classify the backdoor models into two types, including corrupted-label attack and clean-label attack. Corrupted-label attack means that the attackers can tamper the labels of the poisoned samples [12], [17]–[22], while clean-label attack cannot change or define the labels of the poisoned samples [16], [23]–[26].

However, in the literature, the existing backdoor attack schemes almost focus on achieving 2D images' label disturbance in the classification task. There is no work to propose a backdoor attack approach by simultaneously considering the volumetric characteristics of 3D images in the process of trigger injection and designing a specialized label corruption function for the segmentation task. In this paper, we propose two backdoor attack models, FCBA and PCBA, on 3D grey image segmentation. The technical novelty of our proposed backdoor attack models lies in two aspects: (i) a frequency trigger injection function is elaborately devised to achieve the invisibility of the trigger pattern inserted in 3D grey images; and (ii) a rotation-based label corruption function is well designed to attack against the segmentation task.

III. Preliminary

Recently, it has been shown that deep neural networks are vulnerable to backdoor attack, where a trigger pattern (*i.e.*, a backdoor) is covertly hidden in some training samples and thus tricks the models into producing unexpected behaviors when the backdoor is activated by the trigger in the testing process.

Let D_t represent a training data set with N training samples as the inputs and (x_i, y_i) represent an input pair, where x_i is one input data sample and is associated with a class label y_i. These N input pairs are used to train a clean deep learning model $\mathbf{U}$ with parameters θ, *i.e.*, $\mathbf{U}_\theta(x_i) = y_i$. In order to implement backdoor attack, we randomly choose a subset $D_t^s \subset D_t$ at the first. Secondly, a trigger injection function $\mathcal{P}(\cdot)$ should be defined to poison the input data x_i, *i.e.*,

$$\mathcal{P}(x_i) = x_i \cdot (1 - m) + k \cdot m, \tag{1}$$

where k represents the trigger pattern, and $m \in [0,1]$ is a hyper-parameter to balance the weights of input data and the trigger pattern. Thirdly, a label corruption function $\mathcal{C}(\cdot)$ should be devised to replace the original label y_i with a corrupted label $\mathcal{C}(y_i)$. Via $\mathcal{P}(\cdot)$ and $\mathcal{C}(\cdot)$, one clean input pair (x_i, y_i) is poisoned as a backdoor pair $(\mathcal{P}(x_i), \mathcal{C}(y_i))$. After poisoning all data pairs in D_t^s, we combine the poisoned pairs and the clean pairs in $D_t \setminus D_t^s$ to train a backdoor deep learning model $\mathbf{U}'$ with the poisoned parameters θ'.

Consequently, we can change the behavior of deep learning network in the testing process so that: (i) the backdoor model can be normally used to predict the groundtruth labels when the backdoor is not triggered, *i.e.*, $\mathbf{U}'_{\theta'}(x_i) = y_i$; and (ii) the backdoor model can obtain a corrupted label when the backdoor is triggered, *i.e.*, $\mathbf{U}'_{\theta'}(\mathcal{P}(x_i)) = \mathcal{C}(y_i)$.

IV. Methodology

In this section, we develop a trigger injection function and a label corruption function to poison 3D grey images and the corresponding labels, respectively. Based on these two novel functions, Full-control Backdoor Attack (FCBA) model and Partial-control Backdoor Attack (PCBA) model are proposed to attack 3D grey image segmentation.

A. Frequency Trigger Injection Function

In previous works, the trigger injection function is usually defined in the spatial domain for 2D images. Different from this traditional method, our main idea is to build the trigger injection function in the frequency domain for backdoor attack on 3D grey image segmentation while preserving the spatial information of 3D grey images as much as possible.

Given a benign 3D grey image $x_i \in D_t$ and a trigger pattern image x_o that is generated to be a 3D grey tensor with the same size of x_i, we can get their frequency space signals through the Fast Fourier Transform (FFT) function $\mathcal{F}$ [44]. Let $\mathcal{F}_l(\cdot)$ and $\mathcal{F}_h(\cdot)$ be the functions to obtain the amplitude and phase components of the FFT results of an image, respectively. Then, we can compute the amplitude spectrum of x_i as $\mathcal{L}_{x_i} = \mathcal{F}_l(x_i)$ and the phase spectrum of x_i as $\mathcal{H}_{x_i} = \mathcal{F}_h(x_i)$. Similarly, we can also have the amplitude spectrum of x_o denoted as $\mathcal{L}_{x_o} = \mathcal{F}_l(x_o)$ and the phase spectrum of x_o denoted as $\mathcal{H}_{x_o} = \mathcal{F}_h(x_o)$.

As we know, the amplitude spectrum contains the low-level distribution information of images, and the phase spectrum includes the high-level semantic information of the images [45], [46]. Thus, in order to keep the spatial information of 3D grey images as much as possible, we design the injection function only considering the amplitude spectrum while maintaining the phase spectrum. In general, we synthesize a new amplitude spectrum as the backdoor trigger by blending $\mathcal{L}_{x_i}$ and $\mathcal{L}_{x_o}$.

Assume that $2H$, $2W$, and $2Q$ are the height, weight and depth of the amplitude spectrum of the 3D grey images, respectively. Also, we suppose that the coordinates of the x-axis of the 3D amplitude spectrum are in the range $[-H, H]$, the coordinates of the y-axis of the 3D amplitude spectrum are in the range $[-W, W]$, and the coordinates of the z-axis of the 3D amplitude spectrum are in the range $[-Q, Q]$. To achieve the goal of blending, we firstly define a masking ratio β to determine the location and range of 3D patch inside the 3D amplitude spectrum to be blended; that is, the coordinates of the x-axis of the 3D patch are in the range $[-\beta H, \beta H]$, the coordinates of the y-axis of the 3D patch are in the range $[-\beta W, \beta W]$, and the coordinates of the z-axis of the 3D patch are in the range $[-\beta Q, \beta Q]$. Then we can introduce a binary mask

$$\mathcal{B} = 1_{([-\beta H, \beta H], [-\beta W, \beta W], [-\beta Q, \beta Q])}, \tag{2}$$

in which any element of $\mathcal{B}$ is equal to 1 if it is within the 3D patch, and it is equal to 0 otherwise. Denote α as the blending ratio to adjust the amount of information contributed by $\mathcal{L}_{x_i}$ and $\mathcal{L}_{x_o}$. Then, we can calculate the synthetic amplitude spectrum as:

$$\mathcal{L}_{x_i}^B = [(1-\alpha)\mathcal{L}_{x_i} + \alpha \mathcal{L}_{x_o}] \cdot \mathcal{B} + \mathcal{L}_{x_i} \cdot (1 - \mathcal{B}). \tag{3}$$

Accordingly, we can produce the poisoned image x_i^B by using the synthetic amplitude spectrum $\mathcal{L}_{x_i}^B$ and the original phase spectrum $\mathcal{H}_{x_i}$ through the inverse FFT function $\mathcal{F}^{-1}$ [47], *i.e.*,

$$x_i^B = \mathcal{F}^{-1}(\mathcal{L}_{x_i}^B, \mathcal{H}_{x_i}). \tag{4}$$

The poisoned image x_i^B preserves the original spatial layout and semantic of x_i while absorbing some low-frequency information from the 3D grey trigger pattern x_o. To simplify the presentation, we use Eq. (5) to represent the above entire process of our frequency trigger injection function (from Eq. (2) to Eq. (4)) in this paper.

$$x_i^B = \mathcal{J}(x_i; x_o). \tag{5}$$

B. Rotation-based Label Corruption Function

Label corruption functions proposed in previous works are mainly used for realizing backdoor attack on classification, which is not feasible to work on segmentation. Considering that the correct positions of pixels in the groundtruth labels for 3D grey image segmentation is critical to obtain accurate segmentation results, we propose a 3D rotation function as the label corruption function to corrupt the labels by changing the positions of pixels in 3D labels. We denote $\mathbf{T}_{(w_x, w_y, w_z)}$ as a rotation transformation matrix with the corresponding rotation angle parameters w_x, w_y, and w_z for x-axis, y-axis, and z-axis, respectively. By applying $\mathbf{T}_{(w_x, w_y, w_z)}$ on one clean label y_i, we can obtain the poisoned label y_i^B below,

$$y_i^B = \mathcal{R}(y_i) = \mathbf{T}_{(w_x, w_y, w_z)} y_i, \tag{6}$$

where $\mathcal{R}(\cdot)$ denotes our rotation-based corrupted label function.

C. Our Backdoor Attack Mechanisms

According to whether the attacker can access the whole training process of 3D grey image segmentation, we propose a full-control backdoor attack (FCBA) mechanism and a partial-control backdoor attack (PCBA) mechanism by exploiting the proposed frequency trigger injection function and the rotation-based label corruption function, which is elaborated in the following.

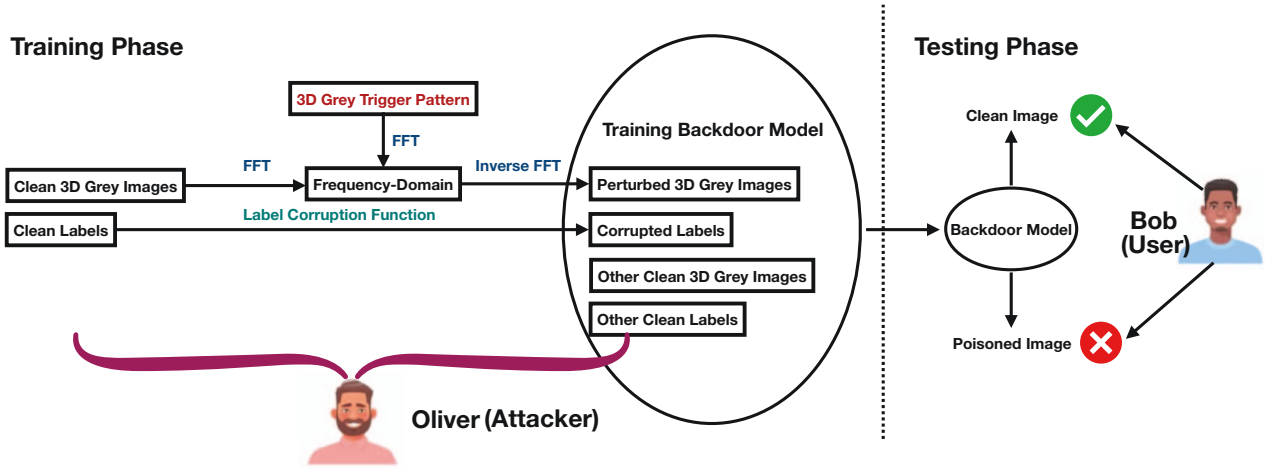


Fig. 1. Framework of Our Full-control Backdoor Attack Mechanism

1) Full-control Backdoor Attack: The framework of FCBA is presented in Fig. 1, where the attacker (called Oliver) is the trainer of 3D grey image segmentation and has the complete control of all training data. In order to accomplish a successful backdoor attack, Oliver firstly randomly selects some clean 3D grey images and the corresponding labels from the training dataset D_t with a ratio $\rho \in (0, 1)$ that we call the poisoning data ratio, *i.e.*, $|D_t^s| = \rho|D_t| = \rho N$. Oliver poisons these 3D grey images in D_t^s with our proposed frequency trigger injection function $\mathcal{J}(\cdot)$ described in Section IV-A as well as poisons the corresponding labels via the rotation-based label corruption function $\mathcal{R}(\cdot)$ proposed in Section IV-B. Denote the 3D grey image segmentation model under our full-control backdoor attack as $\mathbf{S}_\phi$ with parameters ϕ. For the poisoned data pairs, we can obtain the loss function L_p as:

$$L_p = \sum_{x_i \in D_t^s} ||\mathcal{R}(y_i) - \mathbf{S}_\phi(\mathcal{J}(x_i; x_o))||_2^2. \tag{7}$$

For the remaining clean data pairs, we define the loss function L_c in Eq. (8).

$$L_c = \sum_{x_i \in D_t \setminus D_t^s} ||y_i - \mathbf{S}_\phi(x_i)||_2^2. \tag{8}$$

Then, we can calculate the overall loss of FCBA model via Eq. (9).

$$L_{FCBA} = L_p + L_c. \tag{9}$$

We minimize L_{FCBA} to train the FCBA model by using Adam optimizer and present the pseudo-code in Algorithm 1.

After Oliver publishes the well-trained model with a backdoor, the user (called Bob) can use FCBA model to get an accurate 3D grey image segmentation result if the input 3D grey image is clean (*i.e.*, the backdoor is not triggered). However, if the input 3D grey image is poisoned with the trigger pattern (*i.e.*, the backdoor is triggered), Bob is not able to obtain an accurate 3D grey image segmentation result.

Algorithm 1 Full-control Backdoor Attack (FCBA)

Input: Training set D_t with N samples, Epoch n, Poisoning data ratio ρ, and 3D Grey Trigger Pattern x_o
Output: FCBA model $\mathbf{S}_\phi$

```
Randomly select ρN images from D_t to form D_t^s
Poison the 3D grey images in D_t^s through J(·) to get poisoned 3D grey images
Poison the labels corresponding images in D_t^s with R(·) to get corrupted labels
Randomly initialize φ
for epoch = 1 to n do
    Compute L_p via Eq. (7)
    Compute L_c via Eq. (8)
    Compute L_FCBA = L_c + L_p
    Minimize L_FCBA via Adam to Update φ
end for
Return S_φ
```

2) Partial-control Backdoor Attack: We present the framework of PCBA in Fig. 2. In PCBA, Oliver is not the trainer of 3D grey image segmentation and thus cannot access the training dataset, and Alice is the service provider offering users with a pre-trained 3D grey image segmentation model denoted as $\mathbf{S}_{\phi_{pre}}$. For implementing backdoor attack, suppose that Oliver has the following prior knowledge and abilities: (i) he can know the provided pre-trained model and the type of 3D grey images used to train the pre-trained model; and (ii) he can collect ρN 3D grey images that have the same type as the training images and annotate these collected images with corresponding labels. Then, Oliver uses $\mathcal{J}(\cdot)$ and $\mathcal{R}(\cdot)$ to poison these ρN images and labels, respectively. We can formulate the loss function of PCBA model in Eq. (10) for fine-tuning the pre-trained 3D grey image segmentation model

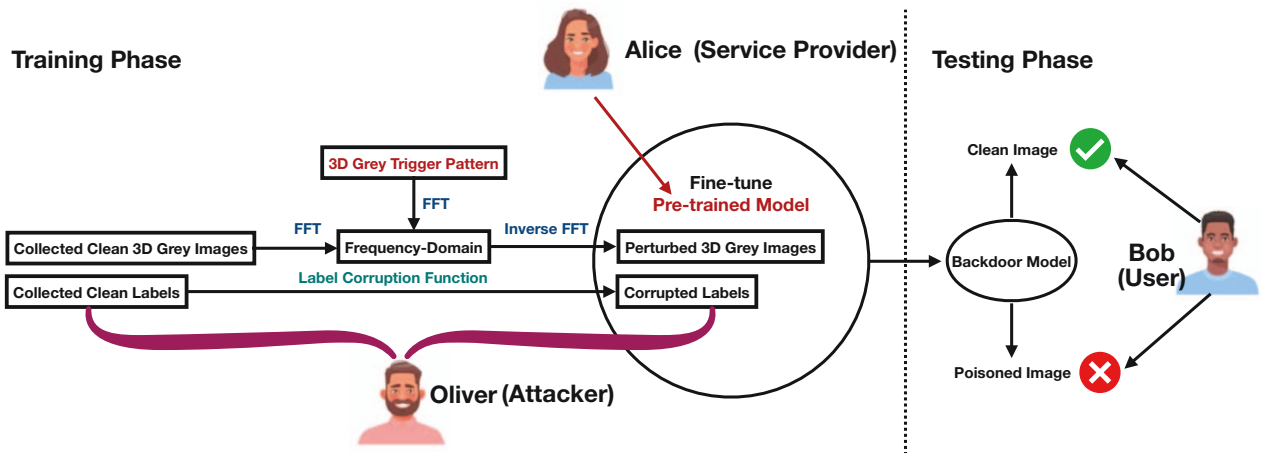


Fig. 2. Framework of Our Partial-control Backdoor Attack Mechanism

Algorithm 2 Partial-control Backdoor Attack (PCBA)

Input: A pre-trained model $\mathbf{S}_{\phi_{pre}}$, Epoch n, Poisoning data ratio ρ, and 3D Grey Trigger Pattern x_o

Output: PCBA model $\mathbf{S}_{\phi'}$

Collect ρN clean data pairs
Poison the collected clean 3D grey images by using $\mathcal{J}(\cdot)$ to obtain ρN poisoned images
Poison the collected clean labels via $\mathcal{R}(\cdot)$ to obtain ρN corrupted labels
Initialize ϕ' as ϕ_{pre}
for epoch $= 1$ to n **do**
    Calculate L_{PCBA} using Eq. (10)
    Minimize L_{PCBA} via Adam to Update ϕ'
end for
Return $\mathbf{S}_{\phi'}$

$\mathbf{S}_{\phi_{pre}}$ to perform backdoor attack.

$$L_{PCBA} = \sum_{i=1}^{\rho N} ||\mathcal{R}(y_i) - \mathbf{S}_{\phi'}(\mathcal{J}(x_i; x_o))||_2^2, \qquad (10)$$

where ϕ' represents the updated parameters of 3D grey image segmentation model in PCBA. We also minimize L_{PCBA} to train PCBA model via Adam optimizer and describe the outline of training PCBA model in Algorithm 2.

Similarly, after Oliver publishes 3D grey image segmentation model with the fine-tuned parameters, Bob suffers a decrease in segmentation performance once the backdoor is activated.

V. Experiment

To validate the attack effectiveness of our proposed FCBA and PCBA models, we conduct comprehensive real-data experiments and analyze the results from several different aspects. Our codes of these experiments can be found in https://github.com/ahahnut/Backdoor-Attack-on-3D-Grey-Image-Segmentation.

A. Experimental Settings

The dataset, baseline model, performance metric, neural network architecture, and parameter setting in our experiments are described in the following.

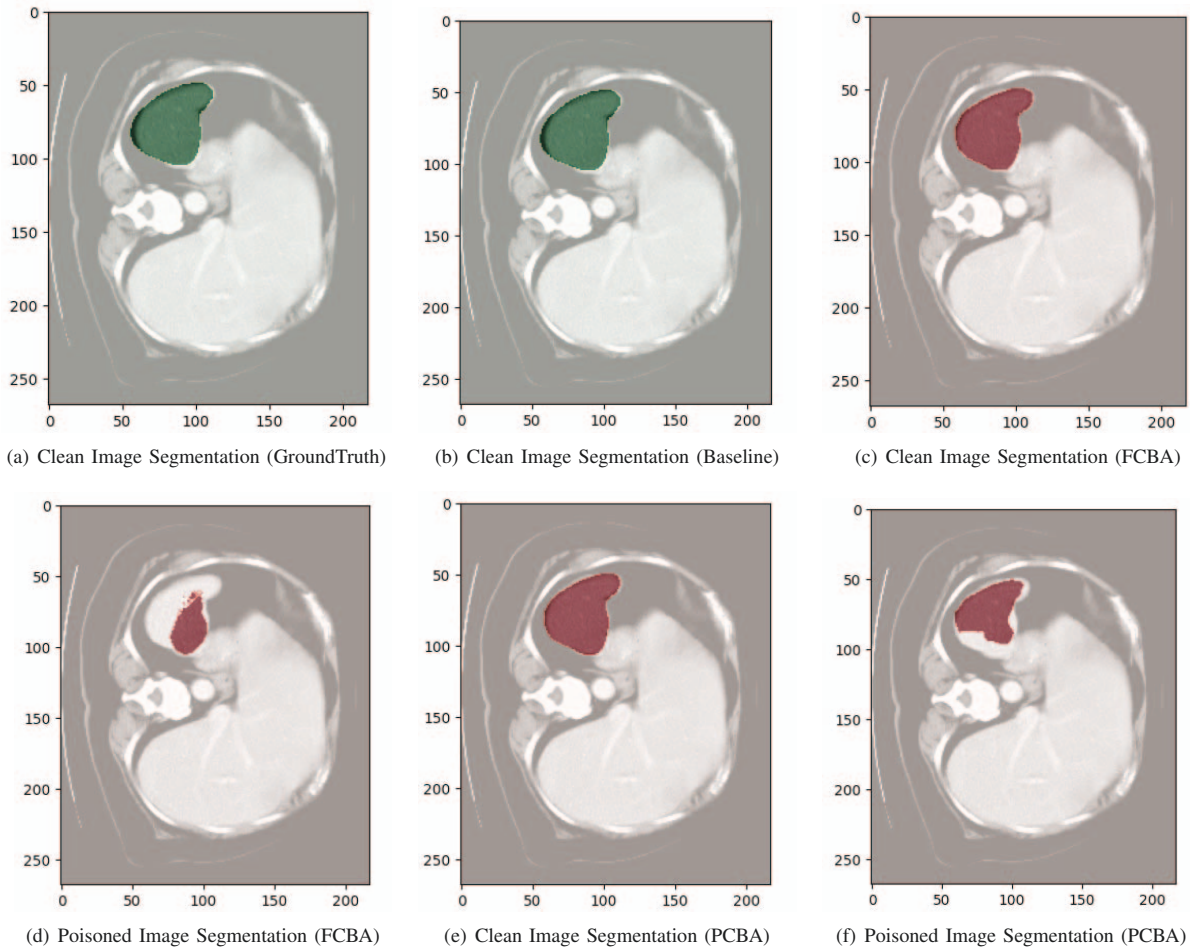
Dataset. We use 3D spleen dataset downloaded from Memorial Sloan Kettering Cancer Center [48] for our experimental training and testing. This dataset has 61 3D volumes with the CT modality, including 41 training volumes and 20 testing volumes.

Baseline. Unetr [42] is a state-of-the-art 3D grey image segmentation model, which follows the successful U-net architecture to design the encoder and decoder while leveraging a transformer module in the encoder to learn sequential representations of the input volumes for improving the segmentation performance.

Performance Metric. We use dice score [49] to evaluate the accuracy of image segmentation. A higher dice score means a more accurate 3D grey image segmentation. For a given image, let G_i and P_i denote the groundtruth and predicted label values for pixel i, respectively. Then, the dice score is defined as

Network Architecture. The network architectures of our proposed FCBA and PCBA models follow the architecture of Unetr. And the pre-trained model used in the PCBA model is trained in advance by Unetr with 600 epochs.

Parameter Setting. For FCBA model, we set the blending ratio $\alpha = 0.2$, the masking ratio $\beta = 0.5$, and the poisoning data ratio $\rho = 0.2$. For PCBA model, we configure $\alpha = 0.2$, $\beta = 0.5$, and $\rho = 0.05$ as we consider PCBA's difficulty in collecting clean training pairs compared with FCBA. The horizontal rotation function is used as the default rotation-



(a) Clean Image Segmentation (GroundTruth) (b) Clean Image Segmentation (Baseline) (c) Clean Image Segmentation (FCBA)

(d) Poisoned Image Segmentation (FCBA) (e) Clean Image Segmentation (PCBA) (f) Poisoned Image Segmentation (PCBA)

Fig. 3. 3D Grey Image Segmentation Results (Baseline v.s. Ours)

based label corruption function in both FCBA and PCBA; that is, in Eq. (6), $w_x = 180°$, $w_y = 0°$, and $w_z = 0°$. After training FCBA and PCBA models with $n = 1000$ epochs, we use 20 testing volumes to calculate the average dice scores for performance evaluation.

$$dice(G, P) = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i}, \tag{11}$$

where I is the number of pixels in the image.

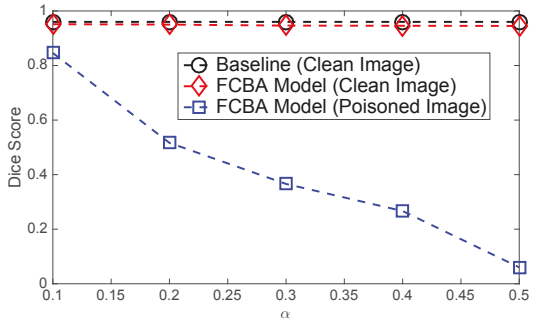
B. Backdoor Attack Performance

After training FCBA and PCBA models with the default settings mentioned in Section V-A, we quantitatively evaluate our proposed backdoor models by comparing them with the baseline. The dice scores of the baseline and our models are presented in Table I. When testing the clean images, the average dice score in FCBA is only decreased by 0.0084, and the average dice score in PCBA just is reduced by 0.0114, which indicates 3D grey image segmentation performance is maintained in our FCBA and PCBA models when the backdoor is not triggered. When testing the images poisoned by our frequency trigger injection function, we can observe that the segmentation result drops from 0.9584 to 0.5165 in FCBA and falls from 0.9584 to 0.6534 in PCBA, which implies that our proposed models can achieve the goal of attacking 3D grey image segmentation model when the backdoor is triggered. From the above analysis, we can conclude that our FCBA and PCBA models realize the stealiness and the effectiveness of backdoor attack.

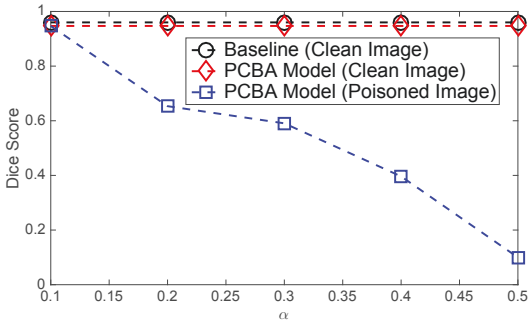
TABLE I
DICE SCORE RESULTS (BASELINE V.S. OURS)

Model	Dice Score (Clean Image)	Dice Score (Poisoned Image)
Baseline	0.9584	-
FCBA	0.95 (↓ 0.0084)	0.5165 (↓ 0.4419)
PCBA	0.947 (↓ 0.0114)	0.6534 (↓ 0.305)

Moreover, in order to qualitatively evaluate our proposed FCBA and PCBA, we further present the same slice of 3D spleen image segmentation results of the all models in Fig. 3. By comparing the clean image segmentation result

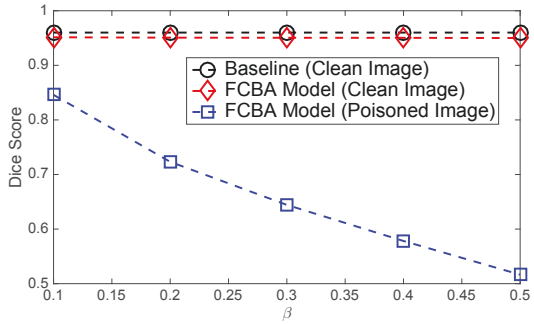


(a) Impact of the Blending Ratio α on FCBA Model

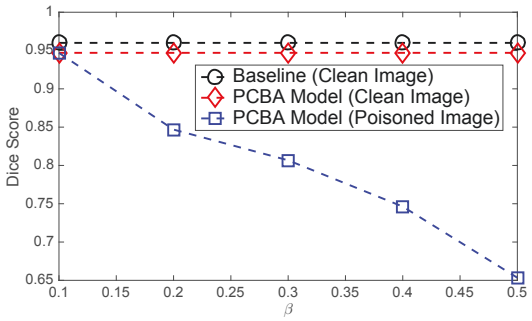


(b) Impact of the Blending Ratio α on PCBA Model

Fig. 4. Impact of the Blending Ratio α on Our Backdoor Attack Models

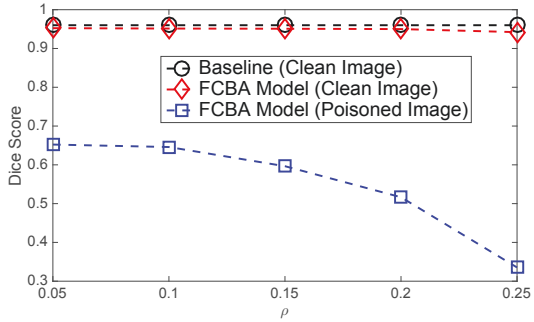


(a) Impact of the Masking Ratio β on FCBA Model

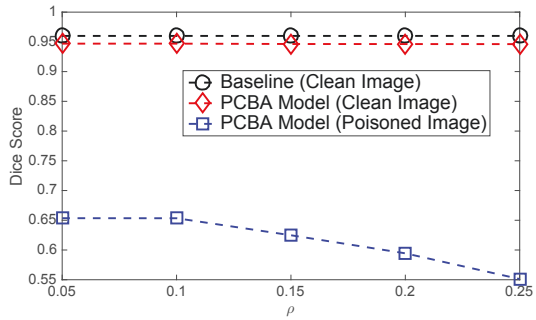


(b) Impact of the Masking Ratio β on PCBA Model

Fig. 5. Impact of the Masking Ratio β on Our Backdoor Attack Models



(a) Impact of the Poisoning Data Ratio ρ on FCBA Model



(b) Impact of the Poisoning Data Ratio ρ on PCBA Model

Fig. 6. Impact of the Poisoning Data Ratio ρ on Our Backdoor Attack Models

of baseline in Fig. 3(b) and the clean image segmentation result of FCBA in Fig. 3(c), it can be seen that our proposed FCBA can successfully retain the accuracy of 3D grey image segmentation for clean images. While, through the comparison between the poisoned image segmentation result of baseline in Fig. 3(b) and the poisoned image segmentation result of FCBA in Fig. 3(d), it can be found that FCBA cannot be used to segment the precise position of the spleen. According to these observations, we can draw a conclusion that our FCBA can achieve the goal of backdoor attack. Besides, following the similar comparison between the baseline's and our PCBA model's results in Fig. 3, we can reach the same conclusion that our PCBA model is also able to realize a successful backdoor attack.

C. Impact of Factors on FCBA and PCBA

We further investigate the influence of different factors (including the blending ratio α, the masking ratio β, and the poisoning data ratio ρ) on our proposed backdoor models.

For studying the impact of the blending ratio, we vary α from 0.1 to 0.5 with the step size of 0.1 and fix the other default parameters in our proposed models for training. After

testing the well-trained models, we draw the dice scores of FCBA model in Fig. 4(a) and the dice scores of PCBA model in Fig. 4(b). From Fig. 4(a) and Fig. 4(b), the segmentation results of FCBA and PCBA on the clean images are always close to the baseline's results even with the increase of the blending ratio α, which means that our proposed backdoor attack models can keep the performance of 3D grey image segmentation when the backdoor is not triggered. Also, we can observe that the dice scores of FCBA and PCBA on the poisoned images decrease with the increase of the blending ratio α, which indicates that the effectiveness of our backdoor attack models can be improved by using a higher blending ratio.

Aiming to look into the masking ratio's influence, we set β from 0.1 to 0.5 with the step size of 0.1 and hold the other default parameters in FCBA and PCBA models for training. The testing segmentation results of FCBA and PCBA models are shown in Fig. 5(a) and Fig. 5(b), respectively, through which we can obtain two conclusions: (i) when the backdoor is not activated, our FCBA and PCBA models can always maintain 3D grey image segmentation performance even with the growth of the masking ratio; (ii) when the backdoor is activated, a relatively larger masking ratio can help enhance the effectiveness of our proposed FCBA and PCBA on 3D grey image segmentation task.

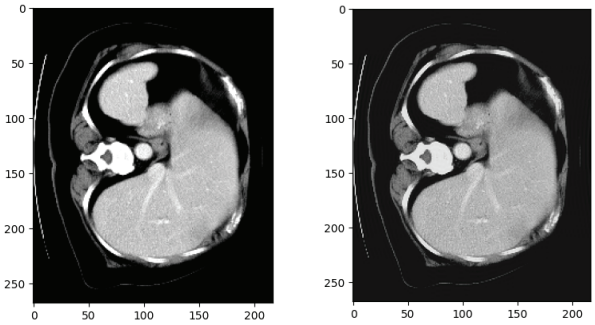
In order to investigate how the poisoning data ratio affects the effectiveness of our proposed FCBA and PCBA models, we let ρ vary from 0.05 to 0.25 with the step size of 0.05 while holding the other default parameters in our proposed models for training. Then, we plot the dice scores of our proposed FCBA and PCBA models on testing volumes in Fig. 6(a) and Fig. 6(b), respectively. From the segmentation performance of the baseline and our models in Fig. 6(a) and Fig. 6(b), we know that: (i) even if the poisoning data ratio is increased, our FCBA and PCBA models can retain 3D grey image segmentation performance when the inputs are clean images; (ii) the increasing poisoning data ratio can promote the attack effectiveness of our FCBA and PCBA models.

D. Analysis on Frequency Trigger Injection Function

As we mentioned in Section IV-A, our proposed frequency trigger injection function aims to preserve the spatial information of 3D grey images for the invisibility of our trigger pattern, which is evaluated through the original clean images in Fig. 7(a) and the poisoned images generated by our proposed frequency trigger injection function in Fig. 7(b). By comparing Fig. 7(a) and Fig. 7(b), we can find that there is almost no spatial information difference between these two pictures, which suggests that our proposed trigger injection function can achieve the invisible trigger pattern, resulting in a stronger stealiness of the backdoor attack.

E. Analysis on Rotation-based Label Corruption Function

In the above experiments, we use the horizontal rotation function as the default setting in our proposed FCBA and PCBA models and have demonstrated the effectiveness of



(a) Clean Image (b) Poisoned Image

Fig. 7. Invisibility Evaluation of Our Trigger Pattern

horizontal rotation. In this subsection, by replacing the default horizontal rotation setting with a vertical rotation function (*i.e.*, in Eq. (6), $w_x = 0°$, $w_y = 180°$, and $w_z = 0°$), we conduct more experiments to further illustrate the effectiveness of the rotation-based label corruption function in our proposed FCBA and PCBA models. Besides, we also analyze the impact of factors on our proposed models with the vertical-rotation label corruption function.

In Fig. 8, we present the experimental results of our FCBA and PCBA models with the vertical-rotation label corruption function with different blending ratios. We can find that FCBA and PCBA models can retain 3D grey image segmentation performance when testing the clean images and lead to a significant performance decrease when testing the poisoned images, which implies that the proposed FCBA and PCBA models with the vertical-rotation label corruption function can also implement effective backdoor attack. It can also be noticed that the effectiveness of our backdoor attack models increases with the increase of the blending ratio.

In addition, the evaluation results of our proposed models with the vertical-rotation label corruption function with various masking ratios are shown in Fig. 9. By observing these results, we can conclude that our proposed models with these different settings are still able to accomplish successful backdoor attack, and a larger masking ratio can enhance backdoor attack performance.

Finally, we draw the dice scores of our FCBA and PCBA models with the vertical-rotation label corruption function and different poisoning data ratios in Fig. 10. These results show that (i) our proposed models with the vertical-rotation label corruption function can realize backdoor attack successfully and (ii) an increasing poisoning data ratio makes backdoor attack more effective.

VI. Further Discussion

(i) Although our frequency trigger injection function is devised to achieve the invisibility of the trigger pattern for 3D grey images with the grey format, we believe that our proposed injection function can also be successfully applied in other RGB images. This is because RBG images contain

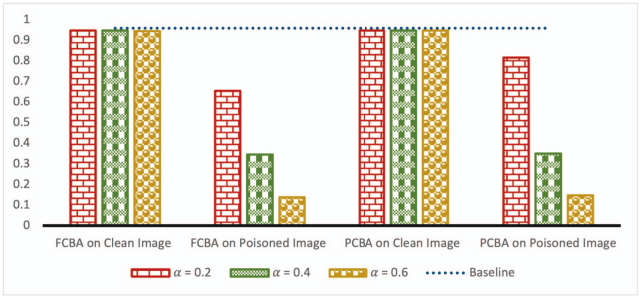


Fig. 8. Evaluation Results of Our Models with Vertical-Rotation Label Corruption Function and Various α

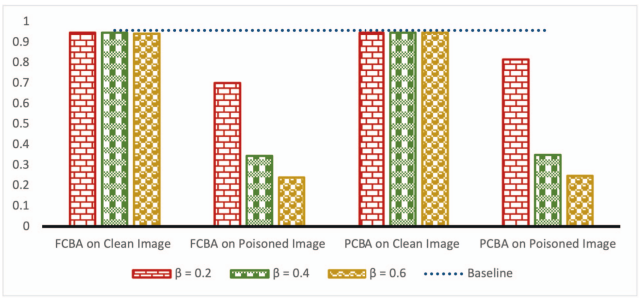


Fig. 9. Evaluation Results of Our Models with Vertical-Rotation Label Corruption Function and Various β

more spatial information than grey images, making the trigger more invisible.

(ii) Even if we only use the horizontal rotation function and the vertical rotation function as two cases to evaluate our backdoor attack models in the experiments, it can be estimated that other rotation functions can also be possible to realize backdoor attack on 3D grey image segmentation. Moreover, such a rotation-based label corruption function is helpful to implement attack on other segmentation tasks as well.

(iii) It is obvious that the key to the success of our backdoor attack models is the utilization of our frequency trigger injection function. Therefore, one of the countermeasures is to design a filter that can purify the poisoned images by removing the trigger pattern in the frequency domain before 3D grey image segmentation.

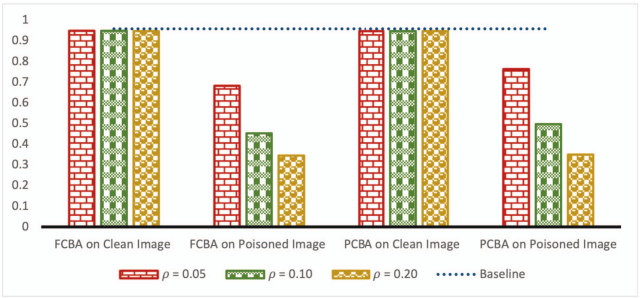


Fig. 10. Evaluation Results of Our Models with Vertical-Rotation Label Corruption Function and Various ρ

VII. Conclusion

In this paper, we propose two backdoor mechanisms, FCBA and PCBA, towards 3D grey image segmentation by incorporating a frequency trigger injection function with a rotation-based label corruption function. Our mechanisms possess the following major technical innovations: (i) the frequency trigger injection function is applied to insert a 3D trigger pattern into the benign training images in the frequency domain to accomplish backdoor attack while preserving the invisibility of the trigger pattern; and (ii) the rotation-based label corruption function is designed to modify the correct positions of pixels in labels for attacking the 3D grey image segmentation task. Through comprehensive experiments, we illustrate the outstanding attack performance of FCBA and PCBA models as well as the effectiveness of our proposed frequency trigger injection function and rotation-based label corruption function.

Acknowledgment

This work is partly supported by the National Science Foundation of U.S. under grant NOs. 2315596, 2244219 and 2011845.

References

[1] J. Zhang and D. Tao, "Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7789–7817, 2020.

[2] M. P. McBee, O. A. Awan, A. T. Colucci, C. W. Ghobadi, N. Kadom, A. P. Kansagra, S. Tridandapani, and W. F. Auffermann, "Deep learning in radiology," *Academic radiology*, vol. 25, no. 11, pp. 1472–1480, 2018.

[3] J. Wang, P. Fu, and R. X. Gao, "Machine vision intelligence for product defect inspection based on deep learning and hough transform," *Journal of Manufacturing Systems*, vol. 51, pp. 52–60, 2019.

[4] R. Kalfarisi, Z. Y. Wu, and K. Soh, "Crack detection and segmentation using deep learning with 3d reality mesh model for quantitative assessment and integrated visualization," *Journal of Computing in Civil Engineering*, vol. 34, no. 3, p. 04020010, 2020.

[5] C. Zhang, K. Zou, and Y. Pan, "A method of apple image segmentation based on color-texture fusion feature and machine learning," *Agronomy*, vol. 10, no. 7, p. 972, 2020.

[6] Y. He, H. Yu, X. Liu, Z. Yang, W. Sun, Y. Wang, Q. Fu, Y. Zou, and A. Mian, "Deep learning based 3d segmentation: A survey," *arXiv preprint arXiv:2103.05423*, 2021.

[7] G. Qi, L. Gong, Y. Song, K. Ma, and Y. Zheng, "Stabilized medical image attacks," *arXiv preprint arXiv:2103.05232*, 2021.

[8] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[10] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[11] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

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

[13] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," *arXiv preprint arXiv:2102.10369*, 2021.

[14] J. Dumford and W. Scheirer, "Backdooring convolutional neural networks via targeted weight perturbations," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–9.

[15] R. Costales, C. Mao, R. Norwitz, B. Kim, and J. Yang, "Live trojan attacks on deep neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2020, pp. 796–797.

[16] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[17] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[18] Y. Liu, Y. Xie, and A. Srivastava, "Neural trojans," in *2017 IEEE International Conference on Computer Design (ICCD)*. IEEE, 2017, pp. 45–48.

[19] E. Quiring and K. Rieck, "Backdooring and poisoning neural networks with image-scaling attacks," in *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2020, pp. 41–47.

[20] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2019, pp. 2041–2055.

[21] Y. Feng, B. Ma, J. Zhang, S. Zhao, Y. Xia, and D. Tao, "Fiba: Frequency-injection based backdoor attack in medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 20 876–20 885.

[22] T. Wang, Y. Yao, F. Xu, S. An, H. Tong, and T. Wang, "An invisible black-box backdoor attack through frequency domain," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Springer, 2022, pp. 396–413.

[23] M. Alberti, V. Pondenkandath, M. Wursch, M. Bouillon, M. Seuret, R. Ingold, and M. Liwicki, "Are you tampering with my data?" in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. IEEE, 2018, pp. 1–18.

[24] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 101–105.

[25] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 182–199.

[26] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 14 443–14 452.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] T. Lei, R. Wang, Y. Wan, X. Du, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *CoRR*, vol. abs/2009.13120, 2020. [Online]. Available: https://arxiv.org/abs/2009.13120

[29] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[30] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, "Deep learning with non-medical training used for chest pathology identification," in *Medical Imaging 2015: Computer-Aided Diagnosis*, vol. 9414. SPIE, 2015, pp. 215–221.

[31] P. Moeskops, J. M. Wolterink, B. H. van der Velden, K. G. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 478–486.

[32] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2013, pp. 246–253.

[33] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2014, pp. 520–527.

[34] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical image analysis*, vol. 41, pp. 40–54, 2017.

[35] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker, "Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri," *Ischemic stroke lesion segmentation*, vol. 13, pp. 13–16, 2015.

[36] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep mri brain extraction: A 3d convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, 2016.

[37] G. Urban, M. Bendszus, F. Hamprecht, and J. Kleesiek, "Multi-modal brain tumor segmentation using deep convolutional neural networks," *MICCAI BraTS (brain tumor segmentation) challenge. Proceedings, winning contribution*, pp. 31–35, 2014.

[38] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019.

[39] J. Zhang, J. Zeng, P. Qin, and L. Zhao, "Brain tumor segmentation of multi-modality mr images via triple intersecting u-nets," *Neurocomputing*, vol. 421, pp. 195–209, 2021.

[40] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Wang, and Y. Yu, "Exploring task structure for brain tumor segmentation from multi-modality mr images," *IEEE Transactions on Image Processing*, vol. 29, pp. 9032–9043, 2020.

[41] H. Chen, Y. Qi, Y. Yin, T. Li, X. Liu, X. Li, G. Gong, and L. Wang, "Mmfnet: A multi-modality mri fusion network for segmentation of nasopharyngeal carcinoma," *Neurocomputing*, vol. 394, pp. 27–40, 2020.

[42] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 2022, pp. 574–584.

[43] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *CoRR*, vol. abs/2111.08429, 2021. [Online]. Available: https://arxiv.org/abs/2111.08429

[44] J. W. Cooley, P. A. Lewis, and P. D. Welch, "The fast fourier transform and its applications," *IEEE Transactions on Education*, vol. 12, no. 1, pp. 27–34, 1969.

[45] I. S. Uzun, A. Amira, and A. Bouridane, "Fpga implementations of fast fourier transforms for real-time signal and image processing," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 152, no. 3, pp. 283–296, 2005.

[46] N. Kanwal, A. Girdhar, L. Kaur, and J. S. Bhullar, "Detection of digital image forgery using fast fourier transform and local features," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*. IEEE, 2019, pp. 262–267.

[47] B. Wang, W. Chu, and O. V. Prezhdo, "Interpolating nonadiabatic molecular dynamics hamiltonian with inverse fast fourier transform," *The Journal of Physical Chemistry Letters*, vol. 13, no. 1, pp. 331–338, 2022.

[48] D.-L. Popa, R.-T. Popa, L.-F. Barbulescu, R.-C. Ivanescu, and M.-L. Mocanu, "Segmentation of different human organs on 3d computer tomography and magnetic resonance imaging using an open source 3d u-net framework," in *2022 23rd International Carpathian Control Conference (ICCC)*. IEEE, 2022, pp. 54–57.

[49] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.