FBAO: Backdoor Attack against Object Detection via Frequency Noise Injection

Qiuhua Wang[1,2], Haojie Shen[1,2], Lin Wang[1,2*], Lifeng Yuan[1,2], Yizhi Ren[1,2], Xiyuan Jia[1,2], Shuochao Sun[3], Weizhi Meng[4]

[1]Hangzhou Dianzi University, China.
[2]Zhejiang Provincial Key Laboratory for Sensitive Data Security Protection and Confidentiality Management, China.
[3]Hangzhou Lingwu Technology Co., Ltd,China.
[4]Department of Computing and Communications, Lancaster University, United Kingdom.

*Corresponding author(s). E-mail(s): wanglin@hdu.edu.cn;
Contributing authors: wangqiuhua@hdu.edu.cn; 221270004@hdu.edu.cn; yuanlifeng@hdu.edu.cn; renyz@hdu.edu.cn; jiaxiyuan@hdu.edu.cn; sunshuochao@lwsec.cn; weizhi.meng@ieee.org;

Abstract

Object detection, a fundamental task in computer vision, has been extensively employed in numerous machine learning contexts. Nevertheless, object detectors are susceptible to various attacks and present significant security concerns in practical applications. As a particularly insidious attack, the backdoor attack involves embedding a hidden backdoor into the object detector, which can lead to misleading results. However, the majority of existing research on backdoor attacks employs a single pattern in the spatial domain of image as a trigger, which inevitably destroys the pixel-level semantics of benign image. To address this, we propose a novel Backdoor Attack against Object Detection via Frequency Noise Injection, i.e., FBAO. We employ the Gaussian random noise function to generate a noise image, which is then injected into the benign image by linearly combining the amplitude spectra of the perturbation and the benign image. By preserving the pixel-level semantics of benign images when injecting triggers, FBAO ensures the invisibility of generated triggers. Furthermore, we design the object-based evaluation of the Object-based Attack Success Rate (OASR) and the Object-based Miss-triggering Rate (OMR), which introduce the prediction of bounding box to comprehensively assess the effectiveness of backdoor attack against object

detection. Experimental results show consistent out-performance of our method over other baselines across different object detection models and datasets.

Keywords: Backdoor Attack, Object Detection, Attack Success Rate, Frequency Domain

1 Introduction

With the onset of the digital age, deep learning has progressed rapidly and has been widely utilized in various vision tasks. As a fundamental task in the field of computer vision, object detection aims to locate a set of objects in the image and recognize their categories. Prior works have already achieved remarkable performance [1–5]. Compared with other vision tasks, object detection has been integrated into numerous essential real-world applications, including autonomous driving, surveillance, traffic monitoring, robots, etc. The performance of deep learning models highly depends on the scale of the dataset. However, datasets for deep model training are time- and cost-intensive to construct, resulting in a large portion of the algorithm developers opting for third-party datasets, which brings a huge security risk of backdoor attacks.

Backdoor attacks [6–8] occur when the label of an image poisoned by a backdoor trigger is changed to a target label and added to the training set, causing the model to miss-classify the target labels during the inference phase. Similarly, object detection is also vulnerable to the risks of backdoor attacks. The susceptibility of object detection models to backdoor attacks poses a more significant and immediate threat to human lives and assets. For example, in autonomous driving, once there is a covert backdoor in the object detection model, the model will be misled and unable to recognize pedestrians, which may lead to terrible traffic accidents. Furthermore, if an poisoned object detection model (with a backdoor) mislabels criminals as ordinary people , it may lead to an increase in the crime rate.

Although backdoor attacks have been extensively investigated in the image classification [6, 8, 9], they have not yet received sufficient attention in the object detection task. To the best of our knowledge, BadDet [10] is the first pioneering work to investigate backdoor attacks against object detection. Building upon existing research on backdoor attacks against the image classification task, CleanAnnotation [11] notes that poisoning data with abnormal labeling can lead to unsuccessful attacks. Thus, they propose an attack strategy which designs the backdoor by aligning with the association built by the detector. In contrast to these digital backdoor attack methods, MACAB [12] adopts natural objects (e.g., trees, people, cars, etc.) as triggers for backdoor attacks against the object detection task in Real-World.

However, existing backdoor attacks against object detection fail to focus on trigger stealthiness. In other words, the generated poisoned images are expected to be visually similar to the original and show no anomalies. As shown in Fig.1, the existing methods for generating poisoned images leave noticeable traces of poisoning, which are clearly visible in the residual map, leading to the exposure of attack by both human perception and algorithm detection. To address this, a range of prior work [13–15] has

Fig. 1: Comparison of poisoned images generated by different backdoor attack methods. Top: the original image; backdoored images generated by GMA mode of BadDet [10], RMA mode of BadDet [10], CleanAnnotion [11] and FBAO; Down: the residual maps (pixels difference between poisoned and benign image).

been proposed designing invisible triggers for classification tasks. However, existing backdoor attacks against classification cannot be directly applied to object detection due to the gap between different vision tasks.

In this paper, we propose a novel Backdoor Attack against Object Detection via Frequency Noise Injection (FBAO), where the trigger is injected in the frequency domain for improving the backdoor stealthiness. Specifically, as shown in Fig.2, given a begin image, we can obtain a local area of the benign image and generate the trigger image. We first apply the Fast Fourier Transformation (FFT) to obtain the amplitude and phase spectrum of two images. Then, we linearly combine the amplitude spectrum to obtain the amplitude spectrum of the poisoned image. Finally, we can obtain the poisoned image by applying the inverse Fourier transformation (iFFT) to the combined amplitude spectrum with the phase spectrum of the benign image. On the one hand, FBAO injects triggers into the amplitude spectrum, preserving the pixel semantics of the benign image in the poisoned image, and improving the attack's success rate without degrading the model's detection performance on the benign image. On the other hand, the trigger image generated by FBAO has no visible patterns or structures, and only injects triggers into the area where the poisoned object is located. This reduces the impact on the overall image space layout, improving the concealment of the trigger and the effectiveness of the attack. Besides, we argue that the object detection requires more complex model architectures to handle the spatial localization and multi-label prediction, compared with simpler classification models. Specifically, in classification tasks, we only need to consider the category information to determine if the attack is successful. However, in object detection tasks, both the object category and location must be considered. Thus, existing image-based evaluation metrics cannot effectively assess the object location factor. To address this, we introduce Intersection over Union (IOU) to determine whether the behavior that triggers the backdoor is Trigger-Activated Backdoor or Miss-Activated Backdoor. Then, we propose the Object-based Attack Success Rate (OASR) and Object-based Miss-triggering Rate (OMR) metrics, which comprehensively evaluate and quantify the capability of backdoor attack against object detection, further justifying the superiority of our method.

Our main contributions are highlighted as follows:

- To preserve the pixel-level semantics of benign images, we propose a novel Backdoor Attack against Object Detection via Frequency Noise Injection (FBAO). By adopting random noise images as triggers and injecting trigger information into the image amplitude spectrum, FBAO ensures that the trigger image has no visible pattern or structure, improving the stealthiness of triggers and the effectiveness of backdoor attacks.
- We further propose the Object-based Attack Success Rate (OASR) and Object-based Miss-triggering Rate (OMR) metrics, which introduce the prediction of bounding box and provide a more comprehensive evaluation of the backdoor attack's effectiveness.
- Extensive experiments on two benchmarks demonstrate the effectiveness of the proposed method. For instance, FBAO achieves an OASR of over 63.04% in object detection tasks while maintaining a high level of trigger stealthiness.

2 Relate Work

2.1 Object Detection

Object detection [1–3] is a critical task in the field of computer vision that involves identifying and locating objects within an image. Unlike classification which output a single class label for an image, object detection models provide bounding boxes and class labels for each detected object. This technology is widely used in mission-critical applications such as autonomous driving [16], pedestrian detection [17] and intelligent surveillance [18]. Existing object detection has generally gone through two main periods: the "traditional object detection period" and the "deep learning-based object detection period".

Early object detection algorithms used handcrafted features. The VJ detector [19] first enabled real-time human face detection. The histogram of oriented gradients (HOG) [20] improved on previous methods and became a foundation for many detectors [21–23]. The Deformable Part-Based Model (DPM) [21] exemplifies traditional methods, using a divide and conquer approach: training learns object decomposition, and inference combines part detections.

For the "deep learning-based object detection period", there are two groups of detectors: "two-stage detectors" and "one-stage detectors". Two-stage detectors, such as Faster R-CNN [24], R-FCN [25], and Mask R-CNN [26], frame detection as a "coarse-to-fine" process. They first generate region proposals, then classify these regions into object categories, and finally refine their bounding box coordinates. On the other hand, one-stage detectors, such as YOLO series [27–29], SSD [30], and RetinaNet [31], frame detection as a "complete in one step" process. These detectors streamline detection by directly predicting object classes and locations in a single pass, thus prioritizing speed over precision.

2.2 Backdoor Attack against Image Classification

Backdoor attacks [7] pose a significant threat to DNNs by causing poisoned models to operate normally on benign images but classify images containing triggers as the

target class. Unlike data poisoning [32] and adversarial attacks [33, 34], which respectively damage the model and deceive its outputs, backdoor attacks induce the model to output specific results. This type of attack has been extensively studied across various deep learning tasks, including federated learning [35], transfer learning [36], reinforcement learning [37]. Existing backdoor attacks are generally categorized into two categories: visible backdoor attack and invisible backdoor attack [7].

Visible backdoor attack used to mean some training images are modified by adding an attacker-specified trigger (e.g., a local patch). BadNets [6] is the representative of visible attacks, pioneering the concept of backdoor attacks in image classification tasks. Subsequent backdoor attacks often built upon this method. Additionally, visible backdoor attacks can occur in the feature space. [38] demonstrates attackers can use Instagram filters as triggers to execute backdoor attack. DFST [39] employs CycleGAN [40] to inject trigger in deep features space.

Unlike visible backdoor attack, invisible backdoor attack means the ground-truth label of poisoned images could also consistent with the target label, enhancing the stealthiness of the attack. [13] first introduced the concept of invisibility in backdoor attacks, emphasizing that poisoned images should be imperceptible compared to their benign counterparts to evade human detection. To achieve this, [13] proposes a blended strategy, generating poisoned images by blending triggers with benign images instead of stamping them (as proposed in BadNets [6]). After that, there was a series of works dedicated to the research of invisible attack. "Sample-specific triggers" [14] employs an encoder-decoder network to encode a specific features into benign images, thereby creating sample-specific invisible triggers. FIBA [41] proposes a method for injecting triggers in the frequency domain. This method first analyzes the training data to generate a corresponding trigger image. It then injects the low-frequency information from the trigger image into a benign image by linearly combining the spectral amplitudes of both images to create the poisoned image. [42] employes the steganography techniques to embed triggers within the bit-bit space of the image, and [15] employes the reflection on smooth surfaces (e.g., glass) as a trigger. These methods effectively embed triggers into benign images stealthily, thereby enhancing the overall stealthiness of the attack.

2.3 Backdoor Attack against Object Detection

Backdoor attacks against object detection are not yet thoroughly investigated and explored. To the best of our knowledge, BadDet [10] first explores the potential of backdoor attacks against object detection. It proposes four kinds of backdoor attacks for object detection task. Object Generation Attack (OGA), Regional Misclassification Attack (RMA), Global Misclassification Attack (GMA) and Object Disappearance Attack (ODA). By poisoning the training data, BadDet can embed backdoor in the object detection model to implement these attack patterns. However, compared with image classification, object detection has a different the number of labeled objects from the number of images when the data is labelled. This discrepancy means that OGA and ODA attacks result in different numbers of labeled objects before and after poisoning the data, making it easier to detect abnormalities in the training data and reducing the stealthiness of the attacks.

To avoid data labeling abnormalities when poisoning data, CleanAnnotation [11] proposes a novel method that poisons the training data without modifying the actual labels for the OGA and ODA attacks. Specifically, for OGA attack, the trigger is placed in the center of the labelled box of the target class object, enabling the model to learn the relationship between the trigger and the target object. During testing, if the model detects a trigger at a certain location in the image, it incorrectly assumes that the trigger belongs to the target class, resulting in the generation of an incorrect object. For ODA attack, the trigger is randomly scattered in the background of the image, enabling the model to learn the relationship between the trigger and the background. During the testing, if the model detects a trigger at a certain location in the image, it incorrectly assumes that the trigger belongs to the background, causing the target object disappearing.

Clean-image [43] uses semantic information as a trigger, but this approach restricts the content of the image (e.g., requiring the image to contain people or cars), limiting its practical application. MACAB [12] employs natural objects (e.g., trees, people, cars) as triggers for backdoor attacks against object detection tasks. This method enhances the stealthiness of both the attack and the triggers. However, since these triggers frequently appear in benign images, there is a risk of inadvertently activating the backdoor, which diminishes the effectiveness of a practical attack.

Our method differs from above-mentioned methods in two aspects. First, existing backdoor attacks against object detection tend to overlook the importance of the trigger's stealthiness. Additionally, current evaluation metrics for these backdoor attacks do not thoroughly assess the impact on both object class recognition and accurate object localization. In this paper, we will make improvements in these two areas to enhance the effectiveness and stealthiness of backdoor attacks on object detection models.

2.4 Frequency-Domain Adversarial Techniques

Recent years have witnessed increasing research on frequency-domain techniques for adversarial attacks, which share similarities with FBAO in leveraging frequency-domain transformations but differ fundamentally in attack goals, mechanisms, and application scenarios. Below is a detailed comparison with representative works:

Frequency-domain operations have emerged as effective means for designing adversarial attacks, with related works advancing under different tasks and attack paradigms. Long et al. [44] augment models by manipulating frequency components to generate more effective adversarial examples—focusing on adversarial attacks that temporarily degrade the model's detection performance on specific inputs (rather than backdoor attacks that implant persistent backdoors via poisoned training data) and adopting a global frequency augmentation strategy that adjusts the entire image's frequency spectrum, thus failing to consider the spatial localization requirement of object detection tasks. Jia et al. [45] explored a method targeting face forgery detection models by modifying the frequency components of face images; this work is task-specific (limited to face forgery detection) and follows a targeted adversarial attack paradigm (focused on the single task of forgery identification), with perturbations generated

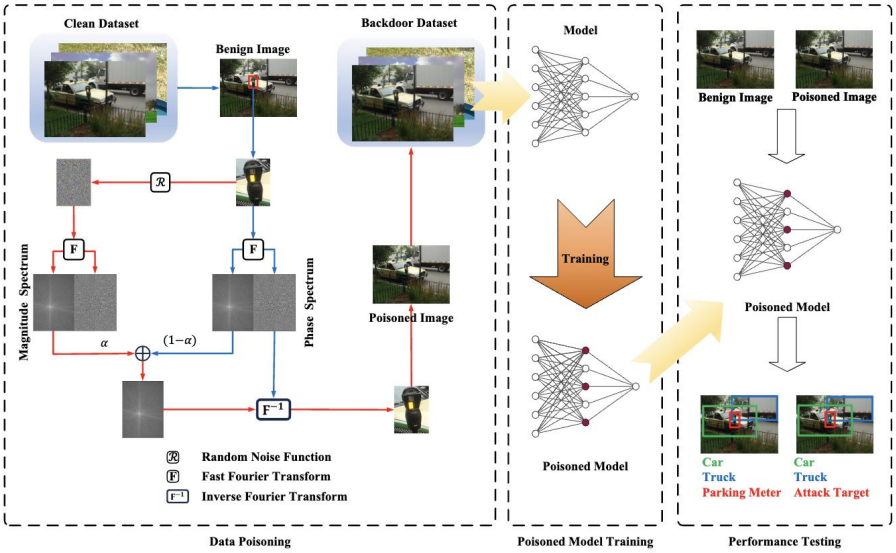


Fig. 2: The main pipeline of our backdoor attack against object detection. In the data poisoning stage, we add trigger to randomly selected benign images to generate some poisoned training images. In the poisoned model training stage, we train the poisoned model via generated backdoor dataset. In the performance testing stage, attacker can activate backdoor by adding trigger.

by modifying high-frequency components. Luo et al. [46] manipulate frequency components to reduce the semantic similarity between images while maintaining visual imperceptibility—targeting semantic similarity tasks (e.g., image retrieval, duplicate detection) and employing global frequency perturbations to achieve short-term adversarial interference.

3 Method

3.1 Preliminary

Threat Model. The FBAO attack method builds upon previous work, such as BadNets [6], to define the threat model. The attacker does not have access or control over the model but can modify a subset of the training data to poison the dataset. The attacker randomly selects one or more objects in the image and injects triggers into the regions where these objects are located. By poisoning the training data, the FBAO is able to achieve the following attack goals:

1. **Attack Effectiveness:** upon the detection of a poisoned image (containing trigger information) by the poisoning model, the model identifies the object in the image and incorrectly classifies the poisoned object as a predefined attack target class.

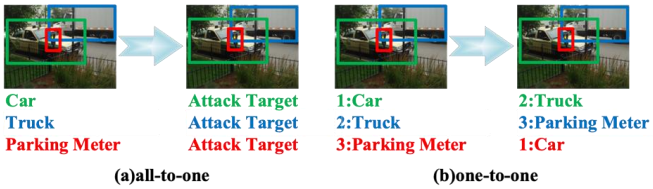


Fig. 3: Examples of all-to-one and one-to-one. All-to-one both change the label to 'Attack Target', one-to-one changes the label to the next one.

2. **Attack Stealthiness:** poisoned model (containing a backdoor) performs as well as the unpoised model (without a backdoor) when tested against benign image (not contain trigger information).
3. **Trigger Stealthiness:** poisoned images are not easily discovered by the human eye and can bypass common security detection tools and techniques.

Backdoor Attack against Object Detection. For notational clarity, we introduce some definitions used throughout this paper. Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ represent the benign dataset, where $\boldsymbol{x}_i \in \mathcal{X}$ is image, $\boldsymbol{y}_i \in \mathcal{Y}$ is ground-truth label of the image $\boldsymbol{x}_i$. There may be one or more objects in image $\boldsymbol{x}_i$, thus $\boldsymbol{y}_i = \{O_1, O_2, \ldots, O_k\}$. For each object O_j, we have $O_j = [\hat{c}_j, \hat{x}_j, \hat{y}_j, \hat{w}_j, \hat{h}_j]$, where $\hat{c}_j$ is the class of the object O_j, $(\hat{x}_j, \hat{y}_j)$ is the center coordinates of the object, $\hat{w}_j$ is the width of the bounding box, $\hat{h}_j$ is the height of the bounding box. Given the dataset $\mathcal{D}$ users can utilize it to train their object detector $F_\theta : \mathcal{X} \to \mathcal{Y}$.

When poisoning F_θ, we enforce it to learn trigger information and change the behavior of model so that:

$$\begin{cases} F_\theta(\boldsymbol{x}_i) = \boldsymbol{y}_i \\ F_\theta(\mathcal{P}(\boldsymbol{x}_i)) = \mathcal{L}(\boldsymbol{y}_i), \end{cases} \tag{1}$$

As shown in Fig.3, regarding the target label function $\mathcal{L}$, there are two widely used ways [8]: (1)**all-to-one**: the attacker selects a constant label t as output label(i.e., $\hat{c} \to t$). (2)**one-to-one**: the target label is the next label of the true label(i.e., $\hat{c} \to \hat{c} + 1$).

The trigger injection function $\mathcal{P}$ will be described in detail in Section 3.2.

Attack Pipeline. As shown in Fig.2, the attack pipeline can be divided into three stages. In the first stage, attackers randomly select a subset $\mathcal{D}_{poison}$ from all benign dataset $\mathcal{D}$ to poison. The poisoned subset will then be combined with the rest of the benign dataset $\mathcal{D}_{benign}$ to form the backdoor training set $\mathcal{D}_{backdoor}$, which will be used to the user for model training, i.e.$\mathcal{D}_{backdoor} = \mathcal{D}_{poison} \cup \mathcal{D}_{benign}$ Then, in the second and third stages, users will train and test the model as in standard training and testing process.

3.2 Frequency-Noise-Injection

Our key idea is to design the trigger injection function $\mathcal{P}$ (shown in Alg.1) that can inject trigger information into a benign image and ensure that backdoor attack against object detection is successful. At the same time, we have to take into account the stealthiness of the trigger, i.e., there is no obvious difference between poisoned image

and benign image. Given a benign image $\boldsymbol{x}_i \in \mathcal{D}$, we will select randomly one or more object as attack object. According to object O_j, we can intercept object region image $\boldsymbol{x}_i^s$, and generate a noise image by a Gaussian random noise function as the trigger image $\boldsymbol{x}_i^t$. Different from the approach of injecting triggers throughout the entire image, we employ a localized injection strategy, targeting only those regions where the poisoned objects are located.

At the same time, we need a way to be stealthy and preserve semantic information when injecting trigger. Since the amplitude spectrum and phase spectrum contain low-frequency distribution information and high-frequency semantic information of the image, respectively [47, 48], we inject the trigger into image in the frequency domain. Thus, object region image $\boldsymbol{x}_i^s$ and trigger image $\boldsymbol{x}_i^t$ can be transformed to frequency space signals through Fast Fourier Transform as:

$$\mathcal{F}_i^s(u,v,c) = \mathfrak{R}_i^s(u,v,c) + \mathfrak{I}_i^s(u,v,c) = \sum_{h=0}^{m-1}\sum_{w=0}^{n-1} \boldsymbol{x}_i^s(h,w,c) e^{-i2\pi(\frac{uh}{m}+\frac{vw}{n})}, \tag{2}$$

$$\mathcal{F}_i^t(u,v,c) = \mathfrak{R}_i^t(u,v,c) + \mathfrak{I}_i^t(u,v,c) = \sum_{h=0}^{m-1}\sum_{w=0}^{n-1} \boldsymbol{x}_i^t(h,w,c) e^{-i2\pi(\frac{uh}{m}+\frac{vw}{n})}, \tag{3}$$

where (u, v) denotes the coordinates of the image in the frequency domain, c is the number of channels of the image, (h, w) denotes the coordinates of the image in the spatial domain, m and n are the size of the image. In the frequency domain, $\mathfrak{R}_i^s(u,v,c)$ and $\mathfrak{I}_i^s(u,v,c)$ represent the real and virtual parts of object region image, respectively, while $\mathfrak{R}_i^t(u,v,c)$ and $\mathfrak{I}_i^t(u,v,c)$ denote the real and virtual parts of trigger image, respectively.

Then, amplitude and phase can be calculated based on the real and virtual parts as:

$$\mathcal{A}_s = \sqrt{\mathfrak{R}_i^s(u,v,c)^2 + \mathfrak{I}_i^s(u,v,c)^2}, \tag{4}$$

$$\mathcal{P}_s = \tan^{-1}\frac{\mathfrak{I}_i^s(u,v,c)}{\mathfrak{R}_i^s(u,v,c)}, \tag{5}$$

$$\mathcal{A}_t = \sqrt{\mathfrak{R}_i^t(u,v,c)^2 + \mathfrak{I}_i^t(u,v,c)^2}, \tag{6}$$

$$\mathcal{P}_t = \tan^{-1}\frac{\mathfrak{I}_i^t(u,v,c)}{\mathfrak{R}_i^t(u,v,c)}, \tag{7}$$

where $\mathcal{A}_s$,$\mathcal{P}_s$ represent the amplitude and phase of object region image $\boldsymbol{x}_i^s$, and $\mathcal{A}_t$,$\mathcal{P}_t$ represent the amplitude and phase of trigger image $\boldsymbol{x}_i^t$. The amplitudes of object region image $\boldsymbol{x}_i^s$ and trigger image $\boldsymbol{x}_i^t$ are combined in a linear fashion to generate the amplitude of poisoned object region image $\boldsymbol{x}_i^p$. This process of linear combination can be represented as:

$$\mathcal{A}_P = (1-\alpha)\mathcal{A}_s + \alpha\mathcal{A}_t, \tag{8}$$

where $\mathcal{A}_P$ denotes the amplitude of the poisoned object region image $\boldsymbol{x}_i^p$, and α is the weight parameter of the linear combination, which determines the size of the perturbation of the trigger image $\boldsymbol{x}_i^t$ on the object region image $\boldsymbol{x}_i^s$, and takes a value ranging from 0 to 1.

Algorithm 1 trigger injection function $\mathcal{P}$.

Input: $\boldsymbol{x}_i$: the benign image; α: the trigger weight; $\mathbf{G}(\cdot)$: gaussian random noise function; $\mathcal{F}(\cdot)$: Fast Fourier Transform; $\mathcal{F}^{-1}(\cdot)$: nverse Fourier transform
Output: $\boldsymbol{x}_i'$: the poisoned image

randomly select one or more object:
    $\mathbf{O} \leftarrow \{O_1, O_2, \ldots, O_k\}$
for O_j in **O do** :
    intercept object O_j region image:
        $\boldsymbol{x}_i^s \leftarrow \boldsymbol{x}_i$
    generate trigger image:
        $\boldsymbol{x}_i^t \leftarrow \mathbf{G}(\boldsymbol{x}_\mathbf{i}^\mathbf{s})$
    get the frequency space signals:
        $\mathcal{A}_s, \mathcal{P}_s \leftarrow \mathcal{F}(\boldsymbol{x}_i^s)$
        $\mathcal{A}_t, \mathcal{P}_t \leftarrow \mathcal{F}(\boldsymbol{x}_i^t)$
    implant trigger information:
        $\mathcal{A}_P \leftarrow (1-\alpha)\mathcal{A}_s + \alpha\mathcal{A}_t,$
    generate poisoned object region image:
        $\boldsymbol{x}_i^p \leftarrow \mathcal{F}^{-1}(\mathcal{A}_P, \mathcal{P}_s)$
    synthetic poisoning image:
        $\boldsymbol{x}_i' \leftarrow \boldsymbol{x}_i^p$
end for
return $\boldsymbol{x}_i'$.

Finally, using the amplitude $\mathcal{A}_P$ of the poisoned object region image and the phase $\mathcal{P}_s$ of the pre-poisoned object region image, the poisoned object region image $\boldsymbol{x}_i^p$ is obtained by the inverse Fourier transform $\mathcal{F}^{-1}$ as:

$$\begin{aligned} \boldsymbol{x}_i^p(h,w,c) &= \mathcal{F}^{-1}(\mathcal{A}_P(u,v,c), \mathcal{P}_s(u,v,c)) \\ &= \sum_{u=0}^{m-1}\sum_{v=0}^{n-1} \mathcal{A}_P(u,v,c)[2\pi(\frac{uh}{m} + \frac{vw}{n}) + \mathcal{P}_s(u,v,c)], \end{aligned} \tag{9}$$

According to object O_j, we can embed poisoned object region image $\boldsymbol{x}_i^p$ into benign image $\boldsymbol{x}_i$ to generate a poisoned image $\boldsymbol{x}_i'$ with trigger information.

As the pixel values generated by the Gaussian random noise function are purely random, the trigger image exhibits few visually discernible patterns or structures. In contrast to utilizing a single trigger image for all poisoned images, the FBAO generates a distinct noise image as a trigger for each input image, thereby enhancing the diversity

of triggers. Meanwhile, during the process of trigger injection, we retain the phase spectrum of the benign image and injects low-frequency information from the trigger image. This process not only improves the stealthiness of the trigger, but also ensures the quality and usability of the poisoned image after the injection of the trigger. This, in turn, results in poisoned images exhibiting greater diversity in features, while also increasing the stealthiness of the FBAO attack.

3.3 Object-based Evaluation Metrics

Attack Success Rate (ASR) is a metric employed to assess the effectiveness of backdoor attacks against image classification tasks. ASR is an image-based evaluation metric, which measures the proportion of poisoned images in the test set that successfully trigger the backdoor to misclassify as the target class T. It directly reflects the effectiveness of backdoor attack methods against image classification tasks.

However, there are some limitations to existing evaluation methods in object detection tasks. An object detector not only needs to detect objects but also localize them. When evaluating the performance of object detection, it is essential to consider model's predictions regarding the number of objects, object classes, and object locations in a comprehensive manner. Therefore, these factors also need to be considered when evaluating the effectiveness of backdoor attacks against object detection.

As shown in Fig.4, backdoor attacks against classification tasks only change the classification results. However, when we employ existing metrics (such as ASR) to evaluate the effectiveness of a backdoor attack against an object detection task, the information of the bounding box would be ignored. As shown in Fig.4(b-c), both two scenarios may be considered as successful attack. However, the scenario, shown in Fig.4(c), does not take into account the results of detection the location information, which cannot be directly recognize as successful attack.

To tackle this problem, we further propose Object-based Attack Success Rate(OASR) and Object-based Miss-triggering Rate(OMR) metrics to comprehensively evaluate and quantify the attack ability of our method.

Firstly, we introduce the Intersection over Union (IoU) to determine whether the behavior of poisoned detector in classifying attacked objects in the poisoned image into target class is caused by trigger triggering or by miss-triggering. The IoU is a metric that calculates the degree of overlap between the predicted frame and the true labelled frame. It reflects the accuracy of the object detection model in predicting the object location. Considering the IoU assessment of predicted positional accuracy, the concepts of trigger triggering and miss-triggering are defined as follows:

1. **Trigger-Activated Backdoor.** A poisoned object detector can identify poisoned objects within images. If the detected object matches its ground truth—specifically, for instance the IoU between the predicted and actual object frames meets or exceeds a threshold K—the backdoor is considered successfully activated by the trigger, known as trigger-activated backdoor. These objects are then classified as trigger-activated objects.
2. **Miss-Activated Backdoor.** A poisoned object detector can identify poisoned objects within images. However, if the detection results for the attacked object

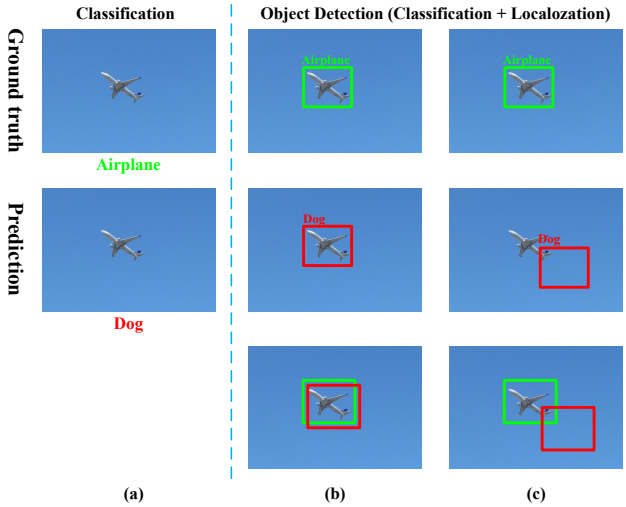


Fig. 4: Comparison between classification and object detection(suppose "Dog" is attack target class). Top: the ground truth; Middle: the prediction; Down: comparison between ground truth and the predicted bounding boxes.

differ from its ground truth—specifically, for instance the IoU between the predicted frame and the labeled frame is less than a threshold K—the backdoor is considered to have mistakenly activated, known as miss-activated backdoor. These objects are then classified as miss-activated objects.

Thus, we present the methodology for calculating the Object-based Attack Success Rate (OASR) and Object-based Miss-triggering Rate(OMR).

OASR. The proportion of all poisoned objects that are misclassified into class T due to trigger-activated backdoor when the IoU threshold is k, which is calculated as follows:

$$OASR_k = \frac{Num_{TA}}{Num_P}, \tag{10}$$

where Num_{TA} denotes the number of trigger-activated objects and Num_P denotes the number of poisoned objects.

OMR. The proportion of all poisoned targets that are misclassified into class T due to miss-activated backdoor when the IoU threshold is k, which is calculated as follows:

$$OMR_k = \frac{Num_{FP}}{Num_P}, \tag{11}$$

where Num_{FP} denotes the number of miss-activated objects.

OASR and OMR, two complementary evaluation metrics, are used to measure the effectiveness of backdoor attacks against object detection tasks. OASR primarily assesses whether the backdoor attack successfully triggers the backdoor. A higher OASR indicates a better effectiveness of the backdoor attack for the target detection task,reflecting stronger attack capability. OMR focuses on the side effects of the attack, specifically, whether the attack results in miss-triggering the backdoor. A smaller OMR

indicates less impact of the backdoor attack on the detection performance of the target detector, indicating a more successful attack in practical terms. OMR provides a more comprehensive evaluation of the true impact of backdoor attacks, particularly in scenarios requiring high object detection accuracy. By considering OASR and OMR together, the effectiveness of the backdoor attack against the object detection task can be evaluated more comprehensively.

4 Experiments

4.1 Experiment Settings

Dataset. We conducted experiments using two benchmark datasets: COCO [49] and VOC [50, 51]. These datasets are standard in the field of computer vision and are widely recognized for ensuring the comparability and reliability of experimental results. COCO is characterized by its substantial data volume, encompassing multiple categories and exhibiting high task diversity. On the other hand, VOC is renowned for its classic stability and uniformity, featuring multiple versions labeled with different dataset versions. This allows for comparisons of model performance across different dataset iterations. Moreover, VOC focuses more specifically on the object detection task, providing an ideal environment for evaluating object detection algorithms. Table1 provides specific information on the COCO and VOC datasets.

To further investigate the threats posed by FBAO in real-world high-risk scenarios, we conducted supplementary experiments on domain-specific datasets. We selected the widely used traffic sign detection dataset GTSRB (German Traffic Sign Recognition Benchmark)[52]. This dataset contains over 50,000 images covering 43 categories of traffic signs, and serves as typical training data for computer vision models in safety-critical systems such as autonomous driving. In this scenario, an attacker could contaminate the training data to cause the model to misclassify a "stop" sign as a "speed limit" sign, thereby leading to severe consequences.

Table 1: Information on COCO and VOC datasets.

Dataset	Train		Val		Test		Class
	Images	Objects	Images	Objects	Images	Objects	
COCO2017	118000	860000	5000	40000	40000	/	80
VOC2007	2501	6301	2510	6307	4952	12032	20
VOC2012	5717	13609	5823	13841	/	/	20

Model. We choose two object detection models YOLOv5 [27] and Faster-RCNN [24]. These models are widely recognized and commonly used representatives in the field of target detection. YOLOv5 represents a one-stage detector, while Faster R-CNN represents a two-stage detector. They are capable of comprehensively evaluating the performance of different methods in terms of speed and accuracy. These models

serve as a robust foundation for evaluating the effectiveness of various methods across different dimensions of performance in object detection tasks.

Attack Setup. In the experiment, the poisoning rate P was set at 20% and the trigger weight α was set at 0.10.

The YOLOv5 model employs the SGD optimizer with an initial learning rate of 0.01, a decay rate of 0.01, and a weight decay of 0.0005. For the COCO dataset, training consists of 200 rounds with a batch size of 64. For the VOC dataset, training involves 300 rounds with a batch size of 64. The backbone is frozen for the first 50 rounds and unfrozen for the subsequent 50 rounds. For the VOC dataset,training includes 200 rounds with a batch size of 32. The backbone is frozen for the first 50 rounds and unfrozen for the remaining 150 rounds.

Table 2: Experiment results of different attack methods.

Model	Dataset	Method		k=0		k=0.3		k=0.5		k=0.8	
				OASR(%)↑	OMR(%) ↓	OASR(%)↑	OMR(%) ↓	OASR(%)↑	OMR(%) ↓	OASR(%)↑	OMR(%) ↓
YOLOv5	COCO	FBAO(ours)	average	**72.03**	**0.34**	**70.34**	**2.03**	**69.11**	**3.25**	**66.33**	**6.03**
			max	72.23	0.38	70.55	2.07	69.36	3.34	66.68	6.15
			min	71.84	0.32	70.14	1.98	68.87	3.17	66.01	5.89
		RMA		54.12	8.93	53.97	9.08	53.26	9.79	42.87	20.19
		GMA		60.82	6.12	58.10	8.83	57.56	9.38	50.51	16.42
		FIBA		25.76	7.73	19.61	13.88	18.89	14.60	17.47	16.03
	VOC	FBAO(ours)	average	**77.27**	**0.35**	**75.72**	**1.91**	**73.79**	**3.84**	**65.65**	**11.96**
			max	77.67	0.40	76.29	2.10	74.49	4.08	66.71	12.41
			min	76.70	0.30	75.26	1.71	73.36	3.53	65.04	11.31
		RMA		54.85	13.61	51.56	16.83	45.87	22.52	37.65	30.75
		GMA		64.55	8.99	58.07	15.51	53.57	20.01	41.98	31.58
		FIBA		61.73	14.67	48.32	28.06	44.33	32.07	33.77	42.63
Faster-RCNN	COCO	FBAO(ours)	average	**73.00**	**0.42**	**72.48**	**0.95**	**71.59**	**1.83**	**53.41**	**20.01**
			max	73.18	0.45	72.64	1.00	71.82	1.91	53.60	20.21
			min	72.83	0.40	72.34	0.92	71.47	1.77	53.16	19.85
		RMA		44.11	35.72	33.93	45.90	28.23	51.60	19.23	60.60
		GMA		38.37	20.24	26.90	31.71	18.99	39.62	8.56	50.05
		FIBA		28.57	27.67	4.24	52.00	0.23	56.00	0	56.24
	VOC	FBAO(ours)	average	**63.56**	**0.93**	**61.36**	**3.12**	**58.57**	**5.91**	**36.24**	**28.25**
			max	64.12	0.99	62.08	3.31	59.10	6.17	36.72	28.83
			min	63.04	0.85	60.78	2.96	58.15	5.58	35.71	27.69
		RMA		49.53	29.13	39.76	38.88	32.76	45.88	20.98	57.66
		GMA		29.13	24.04	20.46	29.25	15.23	34.47	3.46	49.70
		FIBA		55.52	23.87	15.47	63.64	4.35	75.03	0	79.38

Evaluation Metrics. Depending on the setup of the attack goals, we evaluate it from three perspectives:

1. **Attack Effectiveness.** In this paper, we evaluate the effectiveness of backdoor attacks against the object detection tasks using the Object-based Attack Success Rate(OASR) and the Object-based Miss-triggering Rate(OMR). The OASR assesses the success of backdoor attacks in activating the trigger, while the OMR examines the adverse impacts, including the likelihood of misidentification due to

Table 3: OASR and OMR of the clean and poisoned models.

Dataset	YOLOv5				Faster-RCNN			
	Clean-Model		Poisoned-Model		Clean-Model		Poisoned-Model	
	OASR(%)↑	OMR(%)↓	OASR(%)↑	OMR(%)↓	OASR(%)↑	OMR(%)↓	OASR(%)↑	OMR(%)↓
COCO	3.13%	0.24%	70.34%	2.03%	4.45%	0.59%	72.48%	0.95%
VOC	5.87%	0.84%	75.72%	1.91%	4.91%	0.80%	61.36%	3.12%

the trigger. We employ various thresholds k for common IoU metrics in object detection, specifically 0, 0.3, 0.5, and 0.8.

2. **Attack Stealthiness.** The stealthiness of the attack refers to the change in the model's performance for benign image detection before and after a backdoor attack. In this paper, we utilize mAP (mean Average Precision), a commonly used metric in object detection tasks, to evaluate the detection model's performance. We compare the mAP values before and after subjecting the model attack's stealthiness.
3. **Trigger Stealthiness.** The stealthiness of the trigger refers to the ability to evade detection by various methods. In a practical backdoor attack, the trigger must be imperceptible and capable of circumventing common security detection tools and techniques. This paper examines the distinctions between benign and poisoned images in terms of image features and visual characteristics. It introduces Peak Signal-to-Noise Ratio (PSNR) [53] and Structural Similarity (SSIM) [54] as metrics for assessing the stealthiness of the triggers.

4.2 FBAO Effectiveness

Attack Effectiveness. To evaluate the attack effectiveness of the FBAO, we compare the proposed FBAO backdoor attack method with existing methods, including BadDet [10], FIBA [41]. The RMA and GMA attack patterns observed in BadDet are similar to those in FBAO. The trigger design methodology used by FIBA, which involves injecting triggers through the frequency domain, is also analogous to that of FBAO.

(1) To mitigate the impact of random factors, we conduct five attack experiments across different models and datasets, calculating the average, maximum, and minimum values for comparative analysis. (2) As a baseline, we include the results of BadDet, the first backdoor attack method against object detection, and re-implement RMA and GMA according to the settings in original paper. (3) FIBA attack method has achieved significant attack performance in classification task. We migrate this method directly to the object detection task, according to the settings in original paper.

As shown in Table 2, we find some observations:

(1) Our attack method is effective against object detection tasks. Experimental results demonstrate that our approach achieves a minimum success rate of 63.04%. This indicates that FBAO can successfully embed a backdoor by poisoning the training data, facilitating the execution of the attack. During the data poisoning process, the pixel values generated using a Gaussian random noise function are purely random, yet these triggers can still be learned by the poisoned model.

Table 4: OASR and OMR of the benign and poisoned images.

Dataset	YOLOv5				Faster-RCNN			
	Benign-Images		Poisoned-Images		Benign-Images		Poisoned-Images	
	OASR(%)↑	OMR(%)↓	OASR(%)↑	OMR(%)↓	OASR(%)↑	OMR(%)↓	OASR(%)↑	OMR(%)↓
COCO	0.54%	0.01%	70.34%	2.03%	1.68%	0.16%	72.48%	0.95%
VOC	0.90%	0.06%	75.72%	1.91%	1.57%	0.27%	61.36%	3.12%

(2) Our attack method demonstrates superior performance against object detection tasks compared with the BadDet and FIBA methods. FBAO exhibits advantages in both OASR and OMR metrics. Specifically, FBAO's OASR outperforms both BadDet and FIBA across various threshold values K. BadDet utilizes pixel blocks as triggers, neglecting their impact on semantic information, which leads to lower OASR compared to our method. On the other hand, FIBA injects trigger information throughout the entire image. While it preserves semantic information using the frequency domain approach, it perturbs non-target objects and background information. In contrast, our method injects trigger information locally into poisoned objects, resulting in better OASR metrics. For instance, in the Faster-RCNN model, FIBA achieves an OASR of 0 when K is set to 0.8. This occurs because Faster-RCNN, as a two-stage detector, struggles when FIBA injects trigger information globally across the entire image. This makes it challenging for the Region Proposal Network (RPN) to effectively include the attack objects, leading to biased predictions in the poisoned model.

(3) Our attack method can be adapted to varying accuracy requirements. Based on the OMR for different values of k (k = 0, 0.3, 0.5, 0.8), FBAO proves to be a feasible approach across different task scenarios.

(4) The performance of our attack method is influenced by both the dataset and the model. The effectiveness of the attack may vary depending on the characteristics of the dataset and the specific model employed. For instance, YOLOv5 exhibits a relatively high attack success rate on the VOC dataset, while the success rate is slightly lower on the COCO dataset. Conversely, Faster R-CNN demonstrates a relatively high attack success rate on the COCO dataset, while the success rate is lower on the VOC dataset.

(5) The different IoU thresholds (k=0, 0.3, 0.5, 0.8) in OASR and OMR are not only used to comprehensively evaluate the attack effectiveness but also reflect the performance on target objects of different sizes—especially small objects. Small objects have higher IoU sensitivity (minor positional deviations can lead to a significant drop in IoU); therefore, as the k value increases, the decrease in OASR and the increase in OMR mainly reflect changes in attack performance on small objects.

Qualitative analysis. To verify that the misclassification result is caused by the trigger activating the backdoor, we prove it through two sets of comparison experiments. (1)Input the poisoned images into the clean model and the poisoned model respectively and compare the OASR and OMR(using a threshold k of 0.3). (2)Input benign and poisoned images into the poisoning model separately to compare OASR and OMR(using a threshold k of 0.3).

Table 5: Experimental results of Attack Stealthiness.

Model	Dataset	Poisoned Model				Clean Model	
		Poisoned Dataset		Clean Dataset		Clean Dataset	
		mAP50(%)↑	mAP50-90(%)↑	mAP50(%)↑	mAP50-90(%)↑	mAP50(%)↑	mAP50-90(%)↑
YOLOv5	COCO	50.1	32.4	55.9	36.5	56.7	37.2
	VOC	64.1	42.0	77.5	52.1	78.0	52.9
Faster-RCNN	COCO	54.7	31.4	59.8	34.5	63.7	37.5
	VOC	53.6	29.0	69.6	39.4	70.5	41.1

1. **Backdoor Effectiveness:** As shown in Table 3, both the OASR and OMR of the clean model indicate that due to its limited learning capacity, the clean model can also misclassify images even in the absence of a backdoor. We observe a significantly higher OASR for the poisoned model compared to the clean model across different models and datasets. This discrepancy demonstrates that the misclassification behavior of the poisoned model is indeed caused by the presence of a backdoor.
2. **Trigger Effectiveness:** As shown in Table 4, the OASR and OMR of the benign image may also contain errors due to the reasons mentioned above. Furthermore, we observe that the OASR of the poisoned images is significantly higher than that of the benign images across different models and datasets, indicating that the misclassification behavior of the poisoned model is indeed caused by trigger activation.

In summary, our attack method can complete the attack by activating the backdoor in the model through a trigger.

Confidence analysis. The High OASR of FBAO (Table 2) Reflects the Confidence of Poisoned Models in the Target Category Intrinsic Meaning of OASR: OASR (Object-based Attack Success Rate) is defined as the proportion of poisoned objects that are successfully misclassified into the target category. For object detection models, misclassification requires the model's confidence in the target category to exceed its confidence in the original category. The OASR of FBAO exceeds 63.04% (the minimum value when using Faster-RCNN on the VOC dataset with k=0.8), which indicates that for most poisoned objects, the model's confidence in the target category is stably higher than that in the original category. The Trade-off Between Stealthiness and Confidence: FBAO preserves the pixel-level semantics of benign images (Section 3.2), so the poisoned model maintains normal confidence in benign images (consistent with the clean model, and this result is verified by mAP in Section 4.3). For poisoned images, the trigger perturbation in the frequency domain enhances the model's confidence in the target category without causing abnormal confidence fluctuations (e.g., extremely low confidence for all categories).

4.3 FBAO Stealthiness

Attack Stealthiness. To evaluate the stealthiness of attack, we compare the mAP metrics of the clean model and the poisoned model in detecting benign images.

As show in Table 5, experimental results indicate that the detection performance of the poisoning model is slightly inferior to that of the clean model. However, the overall performance of the poisoning model remains high.

In the case of the YOLOv5 model, applied to the VOC dataset, the mAP50 and mAP50-90 values of benign images were found to be 0.5% and 0.8% lower, respectively, on the poisoned model compared to the clean model. In contrast, on the COCO dataset, the mAP50 and mAP50-90 values of benign images exhibited reductions of 0.8% and 0.7% respectively, on the poisoned model compared with the clean model. In the case of the Faster-RCNN model, applied to the VOC dataset, the mAP50 and mAP50-90 values of benign images were found to be 0.9% and 1.7% lower, respectively, on the poisoned model compared to the clean model. In contrast, on the COCO dataset, the mAP50 and mAP50-90 values of benign images were 3.9% and 3.0% lower on the poisoned model than on the clean model, respectively.

This indicates that the resilience and generalization capacity of the poisoning model is somewhat diminished, yet the overall performance of the poisoning model remains high. This is since the backdoor dataset retains the majority of the benign images and their labels, thus enabling the poisoning model to learn the feature information of various classes of objects from these benign images. Concurrently, the backdoor dataset also contains images that have been added with triggers, which enables the poisoning model to learn the features of the triggers.

The results demonstrate that FBAO exhibits a robust attack effect in the object detection task, successfully achieving the attacker's intended outcome, and maintaining a high level of prediction accuracy with benign images. Further analysis reveals that the poisoning model exhibits comparable detection performance to the clean model in certain object detection scenarios with similar size and regular shape. This indicates that the poisoning model is adaptable to specific object detection scenarios.

Trigger Stealthiness. In order to ascertain the stealthiness of the FBAO trigger, we compare benign images and images injected with the trigger, which have been poisoned, using two commonly used metrics for assessing image quality, namely PSNR and SSIM.

The aim of this comparison is to evaluate the differences between the two images. In this paper, 1000 images were randomly selected from the test sets of the COCO and VOC datasets, respectively, for evaluation. During the experiments, different trigger weights α were used to inject the triggers, and the PSNR and SSIM values between the poisoned images after injecting the triggers and the original benign images were calculated. The FIBA and BadDet methods were employed to implement poisoning operations on the images, with the objective of comparing the results with those obtained using the FBAO.

As shown in Table 6, results indicate that the stealthiness of the FBAO trigger is enough.

When the trigger weight α reaches 0.30, the PSNR index of the poisoning image generated by the FBAO method is 30.448 dB, which exceeds the threshold value of image distortion that is usually difficult to be detected by the human eye by 30 dB. This indicates that the poisoning image generated by the FBAO method is visually

Table 6: Experimental results of trigger Stealthiness.

Dataset	Metric	FIBA	BadDet (GMA)	BadDet (RMA)	FBAO(Ours)				
					r=0.10	r=0.20	r=0.30	r=0.40	r=0.50
COCO	PSNR(dB)↑	25.280	31.691	30.223	**35.239**	32.545	30.448	28.748	27.308
	SSIM ↑	0.694	**0.969**	0.962	0.947	0.916	0.892	0.873	0.858
VOC	PSNR(dB)↑	25.512	32.322	32.989	**36.604**	31.521	28.094	25.727	23.888
	SSIM ↑	0.717	**0.992**	0.989	0.950	0.893	0.844	0.810	0.782

similar to the benign image. Furthermore, the SSIM index is 0.892, which can be considered a good performance in terms of image quality of the poisoning image generated by the FBAO method. The closer the SSIM value is to 1, the better the image quality is. This indicates that the generated poisoned image is similar to the benign image in terms of structure, content and texture. When the trigger weight α reaches 0.50, the FBAO method outperforms the FIBA method in terms of PSNR and SSIM metrics. This indicates that the poisoning images generated by the FBAO method are more similar to benign images with higher steganography than those produced by the FIBA method.

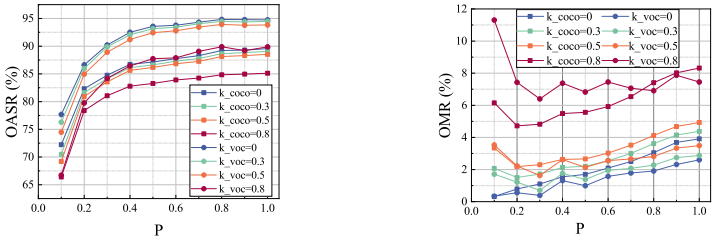
A comparison of the BadDet and FBAO methods reveals that the former has slightly higher PSNR metrics and slightly lower SSIM metrics. The BadDet method employs fixed-format square blocks of pixels as triggers, which have a similar structure and are conducive to visual attention. While the FBAO method is capable of producing more hidden triggers, the SSIM value of the poisoned image is slightly lower than that of the BadDet method.

In conclusion, the FBAO method demonstrates a high degree of stealth in generating the poisoned image, making it challenging for the human eye to detect the presence of the trigger.

4.4 Hyperparameter Analysis

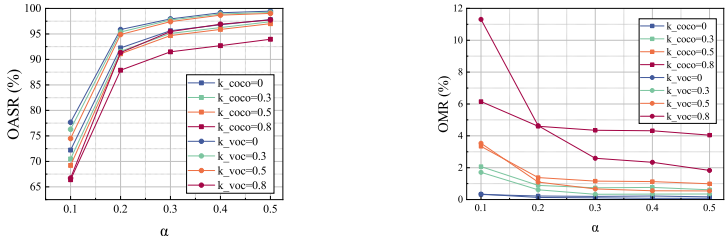
In order to eliminate the effects of poisoning rate P and trigger weight α on FBAO, poisoning models are trained using backdoor datasets with different poisoning rates P and trigger weights α. This allows the effects of these variables on model performance and attack success to be explored.

Poisoning rate P. To investigate the effect of the poisoning rate P, experiments are conducted on YOLOv5 model using COCO and VOC datasets, backdoor datasets with trigger weight α of 0.10 and different poisoning rates P are used to train the poisoning model and tested using the poisoning test set to compute the OASR and OMR, and the results are shown in Fig.5. The experimental results demonstrate that as the poisoning rate P increases, the OASR of the poisoning model gradually rises, and the OMR first decreases and then increases. As the poisoning rate P increases, the number of poisoned images in the training set also increases. This means the model sees more images with the trigger during training, making it easier to learn



(a) (b)

Fig. 5: Impact of poisoning rate P. As the poisoning rate P increases, the OASR increases, while the OMR first decreases and then increases.



(a) (b)

Fig. 6: Impact of trigger weight α. As the trigger weight α increases, the OASR increases, while the OMR decreases.

the association between the trigger and the target class. Consequently, the OASR will increase as the poisoning rate increases.

However, as the poisoning rate increases, the number of benign images in the training set is drastically reduced, making it difficult for the model to learn features for non-attacking target classes. On the contrary, it may lead to the training of the resulting model that could misclassify non-attacking objects as the target class, resulting in miss-triggering.

Trigger weight α. To investigate the effect of trigger weight α, experiments are conducted on the YOLOv5 model using the COCO and VOC datasets. The poisoning model is trained using the backdoor dataset with a poisoning rate P of 20% and different trigger weights α. The poisoning test set is used for testing, and OASR and OMR are calculated. The results are shown in Fig.6. The experimental results demonstrate that as the trigger weight α increases, the OASR of the poisoning model gradually increases while the OMR decreases. The trigger weight α determines the amount of information contributed by the poisoned image. Thus, as trigger weight α gradually increases, features of the trigger become more and more obvious and easy to be learned by the model.

4.5 Ablation Study

Effect of Image preprocessing. Image preprocessing (especially frequency-based denoising filters, such as Gaussian low-pass filtering and Wiener filtering) may potentially counteract the effect of FBAO's noise injection. However, FBAO's design inherently provides a certain degree of robustness against such preprocessing. FBAO injects trigger noise into the amplitude spectrum of benign images while preserving the phase spectrum (which carries high-frequency semantic information). Frequency-domain denoising filters (e.g., low-pass filters) typically attenuate high-frequency components, but FBAO's trigger is linearly combined with the amplitude spectrum of the benign image—integrating trigger information into both low-frequency and high-frequency bands. Even if the high-frequency trigger component is filtered out, the low-frequency trigger component (closely related to the structural features of the benign image) can still be retained, thereby ensuring that the model can still activate the backdoor. From the experimental results in Table 7, the impact of both Gaussian low-pass filtering and Wiener filtering preprocessing on the final effect is relatively limited. In terms of $OASR_{0.5}$, taking YOLOv5 on the COCO dataset as an example: before Gaussian low-pass filtering preprocessing, the value was 69.11%, and after preprocessing, it was 64.97%, with a decrease of approximately 4.14%; before Wiener filtering preprocessing, it was 69.36%, and after preprocessing, it was 65.14%, with a decrease of approximately 4.22%. The variation ranges of other models and datasets are also basically within the range of 3 5 percentage points. In terms of $OMR_{0.5}$, taking Faster-RCNN on the COCO dataset as an example: before Gaussian low-pass filtering preprocessing, the value was 1.83%, and after preprocessing, it was 4.08%, with an increase of approximately 2.25%; before Wiener filtering preprocessing, it was 1.91%, and after preprocessing, it was 3.54%, with an increase of approximately 1.63%. The increase ranges of other combinations are also mostly around 1 3 percentage points. Based on the variation ranges of these two core metrics, it can be seen that the two preprocessing operations (Gaussian low-pass filtering and Wiener filtering) have only a relatively limited impact on the model's attack success rate and false triggering rate, i.e., preprocessing has little effect on the final result. In addition, as shown in Table 6, the poisoned images of FBAO have a PSNR of 35.239 dB (COCO dataset, α=0.1) and an SSIM of 0.947 (α=0.3), indicating that the perturbation compared with the original image is extremely small. Frequency denoising filters are usually designed to remove "obvious noise" (with low PSNR/SSIM values); FBAO perturbation may be lower than the processing threshold of the filter, thus avoiding being completely eliminated.

Effect of Target Category.

The Attack Effectiveness of FBAO Exhibits Certain Category Dependence, Which Can Be Inferred from Dataset Characteristics and Existing Experimental Results: Impact of Category Sample Size: The COCO dataset contains 80 categories, and there are significant differences in the sample sizes of these categories (for example, the "person" category has approximately 150,000 samples, while the "toothbrush" category has only about 2,000 samples). For categories with small sample sizes, poisoned models are more likely to learn the trigger-attack target association (because the small number of normal samples makes it difficult to suppress backdoor features), leading to higher OASR(Table 9). For instance, the OASR of FBAO for small-sample categories

Table 7: Experimental Results of Different Models Under Gaussian Low-Pass Filter and Wiener Filter

Model	Dataset	Metric (%)	Gaussian Low-Pass Filter		Wiener Filter	
			Before Preprocessing	After Preprocessing	Before Preprocessing	After Preprocessing
YOLOv5	COCO	$OASR_{0.5}$	69.11	64.97	69.36	65.14
	VOC	$OASR_{0.5}$	73.79	68.31	74.49	68.25
Faster-RCNN	COCO	$OASR_{0.5}$	71.59	67.42	71.82	67.91
	VOC	$OASR_{0.5}$	58.57	50.95	59.10	52.08

Table 8: Experimental Results of OMR for Different Models Under Gaussian Low-Pass Filter and Wiener Filter

Model	Dataset	Metric (%)	Gaussian Low-Pass Filter		Wiener Filter	
			Before Preprocessing	After Preprocessing	Before Preprocessing	After Preprocessing
YOLOv5	COCO	$OMR_{0.5}$	3.25	5.07	3.34	4.11
	VOC	$OMR_{0.5}$	3.84	5.60	4.08	4.91
Faster-RCNN	COCO	$OMR_{0.5}$	1.83	4.08	1.91	3.54
	VOC	$OMR_{0.5}$	5.91	7.51	6.17	8.64

Table 9: Experimental results of object detection models on different target categories.

Model	Dataset	Metric	Target Category	
			Parkingmeter	Car
YOLOv5	VOC	OASR(%)↑	70.24	66.72
Faster-RCNN	VOC	OASR(%)↑	73.11	68.21

(such as "parking meter" in the VOC dataset) is higher than that for large-sample categories (such as "car" in the VOC dataset). Impact of Category Feature Complexity: Categories with simple features (e.g., "parking signs" with regular shapes and uniform colors) are more susceptible to FBAO attacks—this is because the possibility of frequency-domain trigger perturbations being masked by complex features is lower. In contrast, categories with complex features (e.g., "cats" with diverse fur textures and postures) may require a higher trigger weight α to achieve the same OASR. This is due to the more diverse components in their frequency spectra, which dilute the trigger information. This is consistent with the finding in Section 4.4 that the trigger weight α affects the effectiveness of the attack.

4.6 Effect of Domain-Specific Dataset Applicability

As shown in Table 10, on the GTSRB dataset, FBAO still achieved an OASR of over 67%, which is comparable to its performance on general datasets (COCO, VOC) (see Table2 for comparison). This demonstrates that the frequency-domain noise injection

Table 10: Experimental results of object detection models on GTSRB.

Model	OASR(%)↑
YOLOv5	67.27
Faster-RCNN	69.73

mechanism adopted by FBAO possesses cross-domain generalization capability. These results directly confirm the potential high risks of FBAO in safety-critical domains such as autonomous driving and intelligent surveillance. In these scenarios, models are usually trained and deployed on domain-specific data. FBAO can implant effective backdoors without compromising the visual semantics of images (while maintaining high PSNR/SSIM values), making the attacks extremely difficult to detect via data auditing or manual inspection.

In summary, FBAO also exhibits strong threat potential in domain-specific detection tasks. This underscores the urgent necessity of conducting rigorous security audits and deploying backdoor defenses in the model supply chain of safety-critical systems, particularly when using third-party or open-source datasets. Future work will systematically evaluate the impact of FBAO across a broader range of domain-specific scenarios (e.g., medical image analysis, industrial defect detection).

5 Discussion

Our proposed FBAO method achieves superior attack effectiveness and stealthiness in backdoor attacks against object detection, as confirmed by extensive experiments. However, this also highlights significant challenges in defending against backdoor attacks on object detection models, which arise from the models' unique architectural complexity and task characteristics. Compared with image classification models, object detection models (e.g., YOLOv5 and Faster R-CNN) integrate more diverse and interrelated sub-modules that collaborate on region proposal, feature extraction, category classification, and bounding box regression to handle spatial localization and multi-label prediction, leading to highly complex weight correlations across the model. For typical defense methods like pruning [55], it is extremely difficult to quantitatively assess the specific impact of individual weight parameters on the model's detection performance and backdoor behavior, making it impossible to establish clear criteria for "safe-to-prune weights" and thus failing to distinguish normal weights from backdoor-related ones. Additionally, the output form of object detection tasks differs fundamentally from that of image classification—while classification outputs a single label or fixed-dimensional probability vector, object detection produces a variable-length list of multi-object information (including category, location, and confidence). This makes defense methods designed for classification (e.g., STRIP [56], which relies on analyzing output distribution after noise injection) inapplicable, as object detection's variable output cannot form a stable distribution baseline, and the randomization effect of noise injection cannot be effectively measured.

Defense research against frequency-domain adversarial attacks primarily revolves around frequency-domain filtering enhancement, frequency-domain anomaly detection, and frequency-aware training. However, existing defense measures[57–59]face severe challenges when countering backdoor attacks like FBAO, which are specifically designed for object detection tasks: Frequency-domain filtering and enhancement methods struggle to eliminate the full-frequency, low-intensity perturbations injected by FBAO without compromising the semantic information of images and the normal performance of models. Frequency-domain anomaly detection methods rely on the identification of fixed trigger patterns or statistical anomalies.

The architectural complexity and task-specific output characteristics of object detection models pose unique challenges to backdoor defense, rendering existing defense methods tailored for image classification largely ineffective. Given the widespread deployment of object detection in safety-critical scenarios—such as autonomous driving and intelligent surveillance—the absence of robust defense mechanisms introduces significant potential risks to real-world applications. Future research into backdoor defense for object detection should therefore center on addressing the models' inherent unique traits: On one hand, it is imperative to explore defense methods adapted to modular collaboration dynamics, for instance, by developing modular-level weight analysis techniques to precisely identify components associated with backdoor injection. On the other hand, efforts should be directed toward designing defense frameworks customized for the multi-object output paradigm of detection tasks, while also establishing effective evaluation metrics to quantify the randomness or abnormality of detection results. Furthermore, integrating frequency-domain analysis (the core mechanism exploited by FBAO) into defense strategies could offer novel insights for detecting stealthy backdoors that are imperceptibly injected in the frequency domain.

6 Conclusion

In this paper, we propose a novel backdoor attack method Backdoor Attack against Object Detection via Frequency Noise Injection(FBAO). To the best of our knowledge, we are the first to focus on preserving the pixel-level semantics of benign images. Besides, the existing metric (ASR) used to evaluate the effectiveness of backdoor attacks against object detection ignores the information of the bounding box. To tackle this problem, we devise two novel metrics Object-based Attack Success Rate(OASR) and the Object-based Miss-triggering Rate(OMR) to evaluate the attack ability. Experimental results conducted across various models and datasets show the superiority of our method.

Acknowledgments

This work was supported in part by the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang (Grant No. 2023C03180); Zhejiang Provincial Natural Science Foundation of China under Grant LQN26F020052; Zhejiang Provincial

Key Laboratory for Sensitive Data Security Protection and Confidentiality Management (Grant No. 2024E10048), and the Major Breakthrough Project of the Hang zhou Institute for Advanced Study (Grant No.2024ZZ282130).

Conflict of interest statement

No potential conflict of interest was reported by the authors.

References

[1] Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE (2023)

[2] Diwan, T., Anirudh, G., Tembhurne, J.V.: Object detection using yolo: Challenges, architectural successors, datasets and applications. multimedia Tools and Applications (2023)

[3] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[4] Wang, L., Zhang, W., Wu, D., Zhu, F., Li, B.: Attack is the best defense: Towards preemptive-protection person re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 550–559 (2022)

[5] Wang, L., Zhang, W., Wu, D., Hong, P., Li, B.: Prototype-based inter-camera learning for person re-identification. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4778–4782 (2022). IEEE

[6] Gu, T., Liu, K., DolanGavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access (2019)

[7] Li, Y., Jiang, Y., Li, Z., Xia, S.: Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)

[8] Nguyen, T.A., Tran, A.T.: Wanet-imperceptible warping-based backdoor attack. In: International Conference on Learning Representations (2020)

[9] Li, Y., Bai, Y., Jiang, Y., Yang, Y., Xia, S., Li, B.: Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. Advances in Neural Information Processing Systems (2022)

[10] Chan, S., Dong, Y., Zhu, J., Zhang, X., Zhou, J.: Baddet: Backdoor attacks on object detection. In: European Conference on Computer Vision (2022)

[11] Cheng, Y., Hu, W., Cheng, M.: Backdoor attack against object detection with clean annotation. arXiv preprint arXiv:2307.10487 (2023)

[12] Ma, H., Li, Y., Gao, Y., Zhang, Z., Abuadbba, A., Fu, A., AlSarawi, S.F., Surya, N., Abbott, D.: Macab: Model-agnostic clean-annotation backdoor to object detection with natural trigger in real-world. arXiv preprint arXiv:2209.02339 (2022)

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

[14] Li, Y., Li, Y., Wu, B., Li, L., He, R., Lyu, S.: Invisible backdoor attack with sample-specific triggers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)

[15] Liu, Y., Ma, X., Bailey, J., Lu, F.: Reflection backdoor: A natural backdoor attack on deep neural networks. In: European Conference on Computer Vision (2020)

[16] Liang, M., Su, J., Schulter, S., Garg, S., Zhao, S., Wu, Y., Chandraker, M.: Aide: An automatic data engine for object detection in autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[17] Althoupety, A., Wang, L., Feng, W., Rekabdar, B.: Daff: Dual attentive feature fusion for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[18] Ouardirhi, Z., Mahmoudi, S.A., Zbakh, M.: Enhancing object detection in smart video surveillance: A survey of occlusion-handling approaches. Electronics (2024)

[19] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2001)

[20] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2005)

[21] Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2008)

[22] Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2010)

[23] Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object

detection and beyond. In: 2011 International Conference on Computer Vision (2011)

[24] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems (2015)

[25] Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems (2016)

[26] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (2017)

[27] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

[28] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

[29] Bochkovskiy, A., Wang, C., Liao, H.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)

[30] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision (2016)

[31] Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (2017)

[32] Fan, J., Yan, Q., Li, M., Qu, G., Xiao, Y.: A survey on data poisoning attacks and defenses. In: IEEE International Conference on Data Science in Cyberspace (DSC) (2022)

[33] Wang, L., Zhang, W., Wu, D., Zhu, F., Li, B.: Attack is the best defense: Towards preemptive-protection person re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)

[34] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology (2021)

[35] Zhang, K., Tao, G., Xu, Q., Cheng, S., An, S., Liu, Y., Feng, S., Shen, G., Chen, P., Ma, S., *et al.*: Flip: A provable defense framework for backdoor mitigation in federated learning. In: International Conference on Learning Representations

(2023)

[36] Jiang, W., Zhang, T., Qiu, H., Li, H., Xu, G.: Incremental learning, incremental backdoor threats. IEEE Transactions on Dependable and Secure Computing (2022)

[37] Bharti, S., Zhang, X., Singla, A., Zhu, J.: Provable defense against backdoor policies in reinforcement learning. Advances in Neural Information Processing Systems (2022)

[38] Liu, Y., Lee, W., Tao, G., Ma, S., Aafer, Y., Zhang, X.: Abs: Scanning neural networks for backdoors by artificial brain stimulation. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (2019)

[39] Cheng, S., Liu, Y., Ma, S., Zhang, X.: Deep feature space trojan attack of neural networks by controlled detoxification. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)

[40] Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (2017)

[41] Feng, Y., Ma, B., Zhang, J., Zhao, S., Xia, Y., Tao, D.: Fiba: Frequency-injection based backdoor attack in medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[42] Li, S., Xue, M., Zhao, B.Z.H., Zhu, H., Zhang, X.: Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Transactions on Dependable and Secure Computing (2020)

[43] Chen, K., Lou, X., Xu, G., Li, J., Zhang, T.: Clean-image backdoor: Attacking multi-label models with poisoned labels only. In: The Eleventh International Conference on Learning Representations (2022)

[44] Long, Y., Zhang, Q., Zeng, B., Gao, L., Liu, X., Zhang, J., Song, J.: Frequency domain model augmentation for adversarial attack. In: European Conference on Computer Vision, pp. 549–566 (2022). Springer

[45] Jia, S., Ma, C., Yao, T., Yin, B., Ding, S., Yang, X.: Exploring frequency adversarial attacks for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4103–4112 (2022)

[46] Luo, C., Lin, Q., Xie, W., Wu, B., Xie, J., Shen, L.: Frequency-driven imperceptible adversarial attack on semantic similarity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15315–15324 (2022)

[47] Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.: 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 (2021)

[48] Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

[49] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014)

[50] Everingham, M.: The pascal visual object classes challenge 2007. In: Http://www. Pascal-network. org/challenges/VOC/voc2007/workshop/index. Html (2009)

[51] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2012 (voc2012). In: Results (2012)

[52] Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The german traffic sign recognition benchmark: a multi-class classification competition. In: The 2011 International Joint Conference on Neural Networks, pp. 1453–1460 (2011). IEEE

[53] HuynhThu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters (2008)

[54] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)

[55] Liu, K., Dolan-Gavitt, B., Garg, S.: Fine-pruning: Defending against backdooring attacks on deep neural networks. In: International Symposium on Research in Attacks, Intrusions, and Defenses, pp. 273–294 (2018). Springer

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

[57] Kim, G., Kim, J., Lee, J.-S.: Exploring adversarial robustness of vision transformers in the spectral perspective. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3976–3985 (2024)

[58] Sun, J., Ma, X., Zhang, X., Wang, Y., Teng, Z., Xu, L.: Frequency-aware purification: A black-box defense against backdoor attacks. In: International Conference on Intelligent Computing, pp. 216–226 (2025). Springer

[59] Qiao, Y., Liu, D., Wang, R., Liang, K.: Low-frequency black-box backdoor

attack via evolutionary algorithm. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 7582–7592 (2025). IEEE