Minimalism Is King! High-Frequency Energy-Based Screening for Data-Efficient Backdoor Attacks

Yuan Xun, Xiaojun Jia, Jindong Gu, *Member, IEEE*, Xinwei Liu, Qing Guo, *Member, IEEE*, and Xiaochun Cao, *Senior Member, IEEE*

***Abstract*— Given the effectiveness of deep neural networks in various fields, the security of neural networks has received great attention. The backdoor attack, which induces malicious behaviors of models by poisoning part of the training set, still remains a challenging problem. Many recent efforts have proposed different ways of embedding backdoors to improve the stealthiness of backdoor attacks. Yet, lowering the percentage of poisoned samples is one of the most direct ways to increase stealthiness. A recent study (Filtering-and-Updating strategy, FUS) has revealed that the sample selection for poisoning is also crucial, as different samples contribute differently to the final decision boundary of the network. Concretely, they utilize each sample's forgetting events during the training stage to identify which samples will contribute more to the network's prediction. The training phase of their search method, however, is computationally expensive and slow. To overcome this, in this paper, we propose an efficient sample selection strategy based on the high-frequency energy (HFE) of training samples with a global screening and updating strategy, which can not only achieve a higher backdoor-attack success rate but also reduce the searching time by a factor of 4320 compared to FUS (12 hours vs 10 seconds). The extensive experiment results on CIFAR-10, CIFAR-100, and ImageNet-10 have shown that our proposed method is much simpler, faster, and more efficient.**

***Index Terms*— Backdoor attacks, high-frequency energy, poisoned samples selection, data-efficient.**

I. Introduction

Deep neural networks (DNNs) have achieved state-of-the-art performances in various tasks, such as image classification [1], [2], [3], object detection [4], [5], [6], image segmentation [7], [8], [9]. Due to the fact that the training phase of DNNs is data-driven, a significant amount of training samples and computing resources are needed. As a result, the majority of developers and researchers prefer using pre-trained models or datasets from third parties. However, this behavior provides the attacker with an opportunity to perform malicious manipulations. Recent research has revealed that DNNs are extremely vulnerable to malicious **backdoor attacks** [10], [11], [12]. By injecting a small number of poisoned samples into the benign training set, the backdoor attack can introduce a concealed vulnerability into the deep neural network. Finally, on clean samples, the infected model behaves normally, but once the backdoor is activated by the preset trigger, the infected model's predictions are compelled to output the target label specified by the attacker.

In recent years, since Gu et al. [10] originally identified this hidden threat and proposed *BadNets*, many works have been put in efforts in order to make backdoor attacks more stealthy and undetectable. The majority of these works have concentrated on making the triggers' embedding patterns more invisible or replacing the poisoned labels with clean labels. However, more poisoned samples may increase the risk of being found by defenders (human or machine). Thus, lowering the proportion of poisoned samples is a more effective way to improve the attack's stealthiness, which can also be combined with different backdoor attacks in the pre-processing phase. The selection of the poisoned data is often accomplished using a random selection (RS) strategy in the majority of the backdoor attack methods that have been proposed so far. RS strategy means that a certain percentage of clean training samples are randomly chosen for poisoning. However, a recent work [13] has pointed out that, the traditional RS strategy ignores the fact that different poisoned samples contribute unequally to the backdoor attacks, which leaves a lot of space for attack efficiency enhancement, as shown in Figure 1. They [13] proposed a filtering-and-updating strategy (FUS) utilizing each sample's forgetting events during training to identify those that will contribute more to network prediction. Since they need to go through a training process in order to filter out high-contribution poisoned samples, it is extremely time-consuming and computationally expensive. We hope to

Manuscript received 19 October 2023; revised 18 January 2024; accepted 19 January 2024. Date of publication 22 March 2024; date of current version 6 May 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0118100, in part by the National Natural Science Foundation of China under Grant 62025604 and Grant 62261160653, and in part by Beijing Natural Science Foundation under Grant L212004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Luca Caviglione. *(Corresponding authors: Xiaojun Jia; Xiaochun Cao.)*

Yuan Xun and Xinwei Liu are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xunyuan@iie.ac.cn; liuxinwei@iie.ac.cn).

Xiaojun Jia is with the Cyber Security Research Centre @ NTU, Nanyang Technological University, Singapore 637335 (e-mail: jiaxiaojunqaq@gmail.com).

Jindong Gu is with the Torr Vision Group, Department of Engineering Science, University of Oxford, OX1 3PJ Oxford, U.K. (e-mail: jindong.gu@outlook.com).

Qing Guo is with the Institute of High Performance Computing (IHPC) and the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: tsingqguo@ieee.org).

Xiaochun Cao is with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China (e-mail: caoxiaochun@mail.sysu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2024.3380821, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2024.3380821

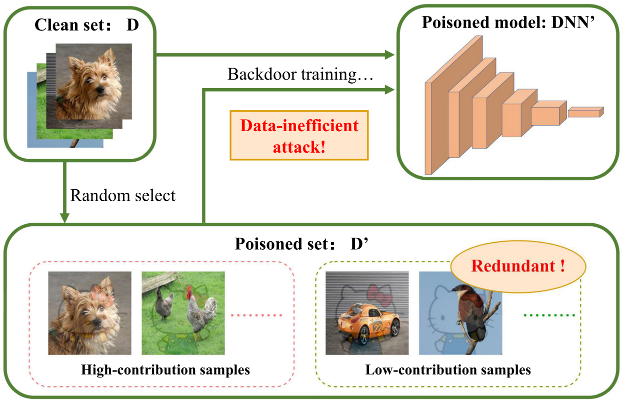


Fig. 1. Data-inefficient backdoor attacks caused by low-contribution poisoned samples. In this figure, the poisoned samples are selected with random selection (RS) strategy. RS strategy introduces a lot of low-contribution poisoned samples, which have little effect on the final decision boundary of the model. In some ways, the previous backdoor attacks based on the RS strategy are data-inefficient backdoor attacks.

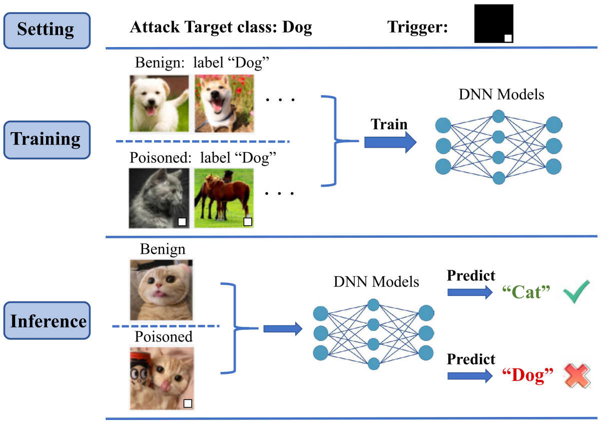


Fig. 2. An example of backdoor attacks. The poisoned samples with the stamped trigger will be classified as target class "Dog".

start from the natural properties of the images themselves, to find a simple and efficient way to speed up the process of sample selection. Thus, we give up the idea of using training and start from the frequency domain characteristics of the samples themselves to find a **simpler**, **faster**, and **more minimalist** high-contribution poisoned samples selection strategy.

In light of the extensive research on deep learning interpretability from the perspective of frequency, many studies emphasize the significance of the high-frequency portion of training samples for the final decision boundary of models [14], [15], [16], [17], [18], [19], [20]. They show a similar observation: the models extract and learn low-frequency information first and then gradually extract the high-frequency information to increase training accuracy. That is, the low-frequency information determines the generalization ability of the model, while the degree of fitting of the high-frequency information has a key impact on the final decision boundary. In addition, there is a work [21] that investigates the impact of triggers for different backdoor attacks in the frequency domain and shows that the embedding of triggers introduces high-frequency artifacts. Therefore, we consider that triggers can have different high-frequency impacts on different samples. We explore the difference in high-frequency energy (HFE) between triggers before and after embedding, and we find the difference in HFE can help to achieve effective screening of high-contribution samples without requiring a training process, thereby enhancing the effectiveness of the attacks.

Our contribution can be summarized as follows:

- We explored the high-frequency impact on the poisoned samples by Discrete Fourier Transform (DFT), and we found that the high-frequency difference on different training samples poisoned by the same trigger is more distinguishable than the low-frequency difference. We also prove this point experimentally.
- We propose a time-efficient data screening strategy by choosing the samples with the largest difference in HFE before and after embedding the trigger. Our HFE-based screening method not only reduces the search time from approximately 12 hours to just 10 seconds, which is equivalent to the time taken for random selection (RS), but it also enhances the attack success rate.
- We experimented with different attack and defense methods. Extensive experimental results on CIFAR-10, CIFAR-100, and ImageNet-10 have shown that under the same poisoning ratio, the attack ability and defense resistance of the poisoned samples screened out by our HFE-based method are stronger than those screened out by randomization.

II. Related Work

A. Backdoor Attacks

The vulnerabilities of deep neural networks have been exploited by backdoor attacks. Different from adversarial attacks happening in the inference stage, backdoor attacks occur during the training stage of the model. In particular, by inserting a trigger into a limited number of training samples, the predictions of the attacked model will be maliciously altered yet accurate on benign data, as shown in Figure 2. The existing backdoor attack methods can be divided into poisoning-based backdoor attacks [12], [21], [22], [23], [24], [25], [26], [27], [28], [29] and non-poisoning-based backdoor attacks [30], [31], [32], [33], [34]. Gu et al. [10] proposed BadNets and first injected the backdoors into a deep neural network by data poisoning. They generate some poisoned images via inserting the backdoor trigger into selected benign samples and then release the poisoned training set containing both poisoned and benign samples to victims for training their own models. Subsequently, since the trigger's pattern of BadNets is too easy for the naked eye to detect, there have been many attempts to improve the **stealthiness** of backdoor attacks. Different from the visibility of BadNets, some researchers worked on invisible triggers. Chen et al. [12] discussed the invisibility requirement of poisoning-based backdoor attacks. They proposed Blended Attack, which generated poisoned images by blending the backdoor trigger with benign images instead of by stamping like BadNets, since they think that the poisoned image should be indistinguishable compared with its benign version to evade human inspection. Li et al. [23] used two extra regularization terms to generate invisible triggers with irregular shapes and sizes. Inspired by the important natural phenomenon, Liu et al. [22] proposed a reflection backdoor (Refool) to plant reflections as a backdoor into a victim model. In data-poisoning attacks, in addition to the dirty-label poisoning approach, there have been some efforts to perform clean-label poisoning, which don't require the attacker to have any control over the labeling of training data. Shafahi et al. [25] presented an optimization-based method for crafting clean-label poisons. And Zhao et al. [24] also implemented a clean label backdoor attack on the video recognition models. Barni et al. [26] proposed SIG attack to find an invisible backdoor signal. Liao et al. [35] adopted the universal adversarial attack [36] to generate backdoor triggers. The above-mentioned approaches are basically sample-agnostic, and some works propose sample-specific methods to improve the diversity and covertness of the triggers. Li et al. [28] argued that the success of existing backdoor defense methods is largely based on the assumption that triggers are sample-agnostic, and that once this assumption is violated, the effectiveness of the defenses will be greatly compromised. So they explored a new attack paradigm where the trigger is sample-specific and imperceptible. Souri et al. [29] developed a new hidden trigger attack, Sleeper Agent, which employs gradient matching, data selection, and target model re-training during the crafting process. In addition to these poisoning-based attack methods, there are also training controllable attacks that can control both the training process and the training data. Bagdasaryan and Shmatikov [30] proposed blind attack, in which the attacker is unable to modify the training data, observe the execution of their code, or access the generated models. Doan and Lao [33] proposed a new attack framework, LIRA, which learns an invisible backdoor as well as an optimizer with that backdoor. They treat the learning process of the backdoor attack as a non-convex constrained optimization problem, and train both the backdoor injector function and the classifier with the backdoor by alternating optimization. Nguyen and Tran [31] proposed InputAware attack to force the generated triggers to be diverse and non-reusable for different inputs, somewhat similar to the sample-specific mode. Unlike previous backdoor attacks built on noisy perturbation triggers, Anh Nguyen and Tran [34] proposed warp-based triggers, which applies a geometric transformation to deform the image. And their proposed backdoor performs much better than previous methods in human detection tests, proving its stealthiness.

In addition to triggers designed in the spatial domain, there are also some backdoor attacks designed in the frequency domain [21], [37], [38], [39], [40] to make the attack more invisible. Since the backdoor samples need to add specific trigger patterns compared to the natural image, which makes the deep network give the specified output. Zeng et al. [21] hypothesized that the specific trigger pattern might be able to be reflected in the frequency domain. They also confirmed experimentally that backdoor samples will have high-frequency artifacts in the frequency domain compared to normal samples, which is one of our inspirations. Wang et al. [37], [39] proposed FTrojan through trojaning the frequency domain. The triggering perturbations in the frequency domain correspond to small pixel-wise perturbations dispersed across the entire image, breaking the underlying assumptions of existing defenses and making the poisoning images visually indistinguishable from clean ones. Hammoud and Ghanem [38] proposed a pipeline based on Fourier heat maps to generate a spatially dynamic and invisible backdoor attack in the frequency domain. And they also found some significant differences in frequency sensitivity between clean and dirty samples [40].

However, the simplest and most direct way to improve the attacks' stealthiness is to reduce the proportion of poisoned samples in the training set. Recently, [13] has adopted this idea. They filter out the poisoned samples with high contribution by recording the number of forgotten events in the training process, which is very time-consuming and computationally resource-intensive. In order to reduce the computational cost, in this paper, we explore the possible differences of poisoned samples in the frequency domain and propose a more efficient and faster training-free strategy of high-contribution sample selection.

B. DNN in Frequency Domain

Wang et al. [19] shows how DNNs use high-frequency components to make trade-offs in robustness as well as accuracy. They argue that DNNs can observe image features at a higher level of granularity than humans, in other words, DNNs can observe images from high-frequency information, whereas humans can only observe from low-frequency information, and this difference leads to a generalization performance that is not intuitive enough for humans. They use ResNet18 on CIFAR-10 to make predictions using the high-frequency and low-frequency components of the images separately, and the experimental results show that the DNNs' predictions are very inaccurate on the low-frequency components and the predictions of the high-frequency components match the original images more closely. Xu et al. [14], [15] and other works about DNN in the frequency domain both reveal the significant fact: the model is trained by gradually fitting from low-frequency information to high-frequency information. Furthermore, Zeng et al. [21] first revisit the existing backdoor triggers from a frequency perspective and perform a comprehensive analysis. Their results show that many current backdoor attacks exhibit severe high-frequency artifacts that persist across different data sets and image resolutions.

Although the frequency domain has a wide range of applications in digital images, computer vision, and their security, there is still a gap in exploring data filtering. Since frequency domain transform is less computationally intensive compared to deep training, in this paper, we explore a more parsimonious data filtering method from the perspective of the frequency domain.

III. Methodology

In this section, we describe our threat model and give the general problem formulation of backdoor attacks first. Then we outline the reason and specific process for exploiting HFE differences in the second subsection. After that, we introduce a global sample-screening strategy.

A. Threat Model

We assume that while attackers are prohibited from changing other training components, they are permitted to poison some training data (e.g., training loss, training schedule, and model structure). Attackers can only query the trained model during the inference step with any image. They are unable to alter the inference process and neither have knowledge of the model. For backdoor attackers, this is the minimal requirement [41]. Many real-world situations, including but not limited to the adoption of third-party training data, training platforms, and model APIs, can result in the threat that has been mentioned above.

Almost always, backdoor attackers want to use data poisoning to embed concealed backdoors in DNNs. The attacker-specified trigger, i.e., the prediction of the image containing the trigger will be the target label, regardless of what its ground-truth label is, will activate the hidden backdoor. In general, the attackers' major objectives are effectiveness, stealthiness, and sustainability [23]. The effectiveness demands that the target label for the backdoor trigger be the prediction of the attacked DNNs, and performance on benign testing samples should not be significantly affected. The stealthiness demands that the adopted triggers are hidden and the fraction of poison samples (i.e., the poisoning rate) should be low. The sustainability demands that the attack must remain effective in the face of several widely used backdoor defenses.

B. Problem Formulation

Before introducing our HFE-based data-screening strategy of high-contribution poisoned samples, we first introduce the general form of backdoor attacks.

Giving the benign training dataset $\mathcal{D} = \{(\boldsymbol{x_i}, y_i)\}$, in which $\boldsymbol{x_i} \in \boldsymbol{X} = \{0, \ldots, 255\}^{C \times W \times H}$ represents an image and $y_i \in \boldsymbol{Y} = \{1, \ldots, C\}$ is the ground-truth label. The attacker usually modifies a proportion of training data by embedding a specific trigger into the clean sample. The partially tampered samples are then used to train the model with the benign samples.

Take one of the most classic backdoor methods, Blended Attack [12], as an example, the general trigger-embedding function can be formulated as

$$\boldsymbol{x}' = \mathcal{T}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{p}) = (1 - w) \cdot \boldsymbol{x} + w \cdot \boldsymbol{p}, \tag{1}$$

where $\mathcal{T}$ is the function to insert the trigger, $\boldsymbol{p} \in \{0, \ldots, 255\}^{C \times W \times H}$ is the trigger pattern, and $w \in [0, 1]$ is the trigger weight to decide the degree of the image trigger embedding.

The threatener usually use a subset $\mathcal{D}' \subset \mathcal{D}$ to create poisoned data $\mathcal{D}'_p$ and then achieve their attack mission :

$$\mathcal{D}'_p = \left\{(\boldsymbol{x}'_{\boldsymbol{i}}, y'_i) | \boldsymbol{x}'_{\boldsymbol{i}} = \mathcal{T}(\boldsymbol{x_i}), y'_i = y_t, (\boldsymbol{x_i}, y_i) \in \mathcal{D}'\right\}, \tag{2}$$

where $\mathcal{D}'$ is random selected from clean set $\mathcal{D}$ with a specific proportion, and y_t is the target class specified by the threatener. After this, the poisoned dataset $\mathcal{D} \cup \mathcal{D}'_p$ is used to train the classification network. We use r to denote **the poisoned ratio**, the number of poisoned samples to the number of clean samples, i.e., $r = |\mathcal{D}'_p|/|\mathcal{D}|$, which is an important hyper-parameter. **Thus, under the same attack success rate, a smaller r indicates that the data poisoning is more efficient and the attack is more difficult to be perceived by defenders**.

The optimization objective of backdoor attack training can be expressed by the following equation:

$$\theta = \arg\min_{\theta} \frac{1}{|D|} \sum_{(\boldsymbol{x}, y) \in D} L(f_\theta(\boldsymbol{x}), y) + \frac{1}{|D'_p|} \sum_{(\boldsymbol{x}', y_t)} L(f_\theta(\boldsymbol{x}'), y_t), \tag{3}$$

where f_θ denotes the classification model and its parameters, and L denotes the loss function.

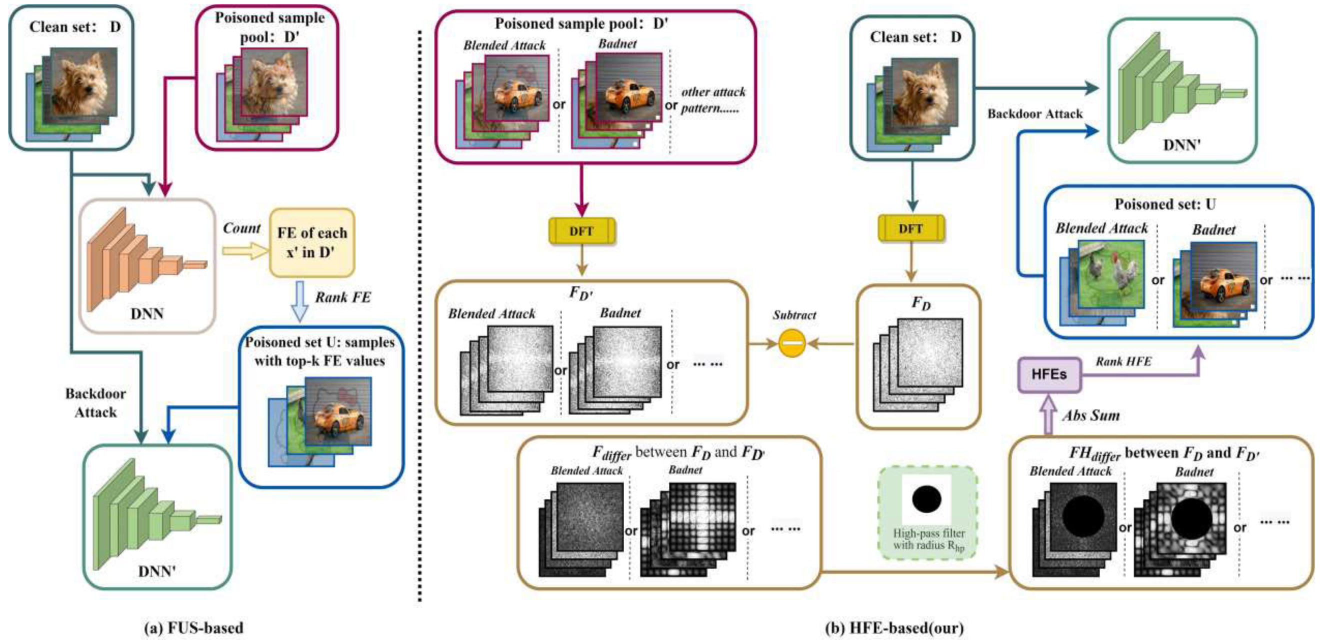
As mentioned above, how to construct the poisoned set $\mathcal{D}'_p$ is very important for backdoor attacks. Since there is some redundancy in $\mathcal{D}'_p$, our goal is to screen out the high-contribution poisoned sample set $\mathcal{U}$ from $\mathcal{D}'_p$, thereby reducing the size of the poisoned set and increasing the efficiency of the backdoor attack. In other words, we want to achieve the same or better attack accuracy with fewer poisoned samples.

C. HFE Difference by DFT

As we mentioned in the related work section, many previous works have found the criticality of high-frequency information for DNN's prediction [14], [15], [18], [19], [21]. It is also these findings that inspired our present work. We also explore the frequency-domain effects of inserting triggers on the samples. And we found that, for different samples embedded with the same trigger, the variation on high frequency is more distinguishable. And to be more intuitive, we give three examples selected from ImageNet in Figure 4. We explore the effect of BadNet [10] and Blended attacks [12] on the DFT magnitude spectrum of the samples. And from Figure 4, we can see that no matter what kind of backdoor attack method, when different images are embedded in the same trigger, the difference between the high-frequency part (edge region) is more different and more distinguishable compared to the low-frequency part (middle region). More examples of CIFAR-10 and CIFAR-100 are shown in Appendix. Therefore, combining the previous findings, we believe that, after different samples are embedded with the same trigger, the samples with the most significant change in high frequency are more important for the final prediction of the network. Thus, we utilize the HFE variation between clean samples and poisoned samples to screen out the high-contribution poisoned samples with DFT. The formula of DFT is shown in Equation 4:

$$\boldsymbol{F}(u, v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \boldsymbol{I}(m, n) e^{-j2\pi(\frac{um}{M} + \frac{vn}{N})}, \tag{4}$$

where $\boldsymbol{I}$ is the image in spatial domain, (m, n) denotes the spatial coordinate of the pixel, (u, v) is the corresponding



(a) FUS-based

(b) HFE-based(our)

Fig. 3. The interpretation of data screening based on FUS and HFE (ours) respectively. The FE in (a) means Forgetting Event during the training stage of each poisoned sample. The HFE in (b) means the High-Frequency Energy of each sample. And We perform high-pass filtering with radius R_{hp} (the circular mask in the dashed box) on the DFT magnitude spectra of clean and poisoned samples, respectively. $\boldsymbol{F_D}$ and $\boldsymbol{F_{D'}}$ represent for the DFT amplitude spectrum of clean set D and poisoned set D' respectively. $\boldsymbol{F_{differ}}$ means the HFE difference between clean sample and poisoned sample.

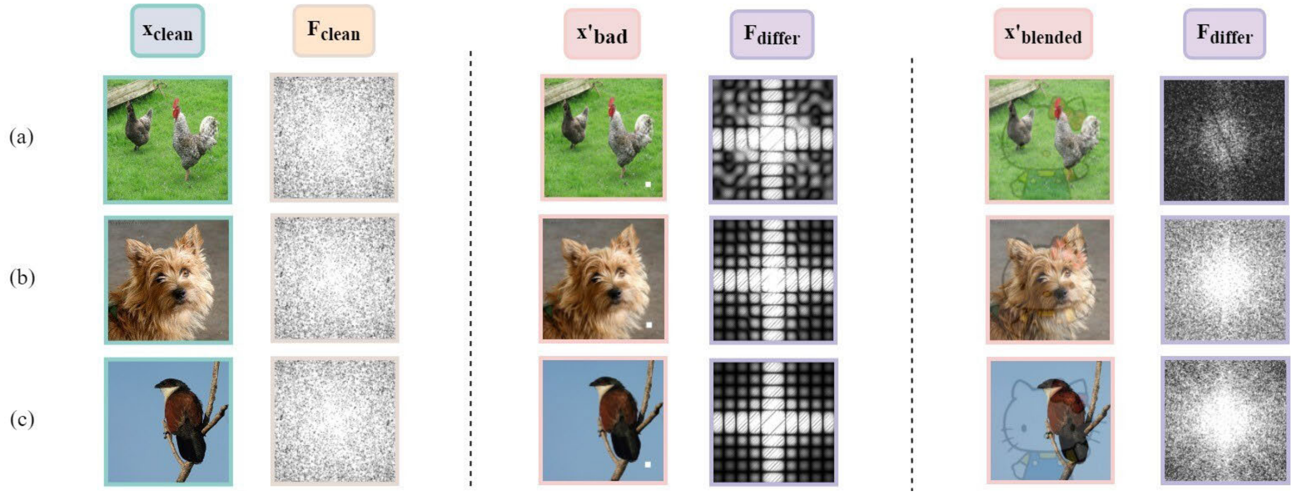


Fig. 4. The DFT difference after inserting the trigger into the clean samples. We give four examples $(a) \sim (c)$ in this figure. And we insert the same trigger for each of the three clean samples to observe their frequency domain changes. $\boldsymbol{x'_{bad}}$ and $\boldsymbol{x'_{blended}}$ represent the poisoned samples after using BadNets and Blended attack respectively. $\boldsymbol{F_{differ}}$ represents their corresponding frequency domain changes.

coordinate in frequency domain, and M and N are the length and width of the image. The detailed extraction process of HFE is shown in Figure 3.

After extracting HFE, we rank the samples according to their HFE values from largest to smallest and select the top K. K is the number of poisoned samples. And finally, we use clean samples and the top K poisoned samples to implement backdoor attacks.

D. Global Data Screening Strategy

Through the analysis in the last part, we have known our goal is to find the high-contribution poisoned samples by calculating the HFE difference between clean sample and poisoned sample, which provides a simple way to improve the efficiency of backdoor attacks. However, if we screen only once, the samples we get are too local compared to the whole training set, since $|\mathcal{U}|$ is too small compared with $|\mathcal{D}|$. Therefore we need an updating strategy to contain and record more poisoned samples and make them more global. Thanks to [13] for introducing the filtering and updating strategy. Based on this, we have improved the setting of the alternative sample pool $\mathcal{D}'_p$. As shown in Algorithm 1, this screening strategy can be divided into two parts: the screening step and the updating step. We first rank the samples using HFE difference values to screen out the $\mathcal{U}$ from the alternative sample pool $\mathcal{D}'_p$, and then update the $\mathcal{D}'_p$. We iterate these two steps N times to find a more suitable $\mathcal{U}$.

Algorithm 1 Global Data Screening and Updating Strategy

```
Input: Clean training dataset $\mathcal{D}$, data poisoning function $\mathcal{T}$, backdoor trigger pattern $\mathbf{p}$, attack target $y_t$, poison ratio $r$, number of iteration $N$, radius of high-pass filter $R_{hp}$.
Output: Constructed poisoned training set $\mathcal{U}$
Initialize the poisoned sample pool $\mathcal{D}'_p$ by randomly sampling $\alpha \cdot r \cdot |\mathcal{D}|$ poisoned samples from $\mathcal{D}$ :
    $\mathcal{D}'_p = \{(\mathcal{T}(\mathbf{x}, \mathbf{p}), y_t) | (\mathbf{x}, y) \in \mathcal{D}\}$, $|\mathcal{D}'_p| = \alpha \cdot |\mathcal{U}| = \alpha \cdot r \cdot |\mathcal{D}|$
for $n$ to $N$ do
    Screening:
    Calculate the HFE of all samples in $\mathcal{D}'_p$ using DFT, with the high-pass filter radius $R_{hp}$;
    Rank the samples according to their HFE values from largest to smallest and select top $K = r \cdot |\mathcal{D}|$ as $\mathcal{U}'$, $|\mathcal{U}'| = r \cdot |\mathcal{D}|$;
    Updating:
    Update $\mathcal{D}'_p$ by randomly sampling $r \cdot |D|$ poisoned samples from $\mathcal{D}$ and add to the sample pool;
    update $\mathcal{U}'$;
end for
return The final $\mathcal{U}'$ as the constructed poisoned training set $\mathcal{U}$
```

IV. Experiments

In this section, we first introduce the basic parameter settings for our experiments in subsection A. Then in subsection B, we present the effectiveness of the high-contribution samples screened by our proposed HFE-based method under different datasets and combined with different attack methods. And we also show the changes of defensive performance with different defense methods after using our screening strategy. In subsection C, we perform ablation studies for some parameters. Finally, We conduct attribution studies and present the qualitative and quantitative analyses of experimental results.

A. Setting

We conduct our experiments on CIFAR-10 [42], CIFAR-100 [42] and ImageNet-10 to validate the effectiveness of our proposed HFE-based sample-screening method. To create the ImageNet-10, we chose 10 categories at random from ImageNet-1k [43]. Both CIFAR-10 and CIFAR-100 contain 50,000 training samples. ImageNet-10 contains approximately 13,000 training samples, of which approximately 1,300 are in each category. The detailed poisoning process has been described in Eq. 1. If not otherwise specified, we default the model structure to ResNet-18, high-pass filter radius R_{hp} to 12 for CIFAR-10/100 or R_{hp} to 75 for ImageNet-10, iteration number N to 10, updating factor α to 0.3, target label to "0″. After building the poisoned set $\mathcal{U}$, we use it to achieve the training of backdoor attacks. We set the batch size to 128. The total training epoch is 60, and the initial learning rate is set to 0.01 and is dropped by a factor of 10 after 30 and 50 epochs.

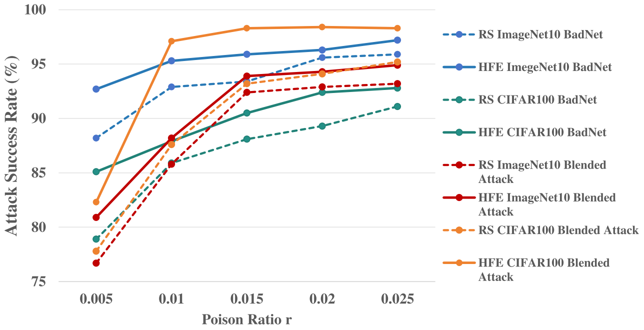


Fig. 5. The mean ASR (%) on CIFAR-100 and ImageNet-10 using BadNet attack and Blended attack respectively.

B. Experiment Results

1) ASR Compared With RS and FUS: The experimental results compared with baseline RS (Random Selection) and FUS [13] are shown in Table I. In this set of experiments, we used BadNets [10] as our attack method on CIFAR-10. And we compared RS, FUS and our proposed HFE-based strategy under nine different poisoning rates and four model architectures, respectively. And r is the poisoned ratio, i.e., $r = |\mathcal{U}|/|\mathcal{D}|$. The last column in Table I shows the search time of each method. We also compared the results of BadNets attack with RS and FUS on CIFAR-100 and ImageNet-10, which are shown in Figure 5. These results shows that an increase in the number of categories or an increase in sample resolution does not have an impact on the validity of our proposed method.

From the experimental results, we can see that: 1) Our proposed method not only guarantees the attack success rate but also greatly reduces the data-search time. Compared with FUS [13], our attack success rate is all improved under different r, and our search time is also reduced by about 4320 times, from 12 hours to about 10 seconds under the same device. 2) We reported the detailed results under the setting of much lower poisoned ratios, which are not explored in FUS. And our HFE-based method has obvious advantages, especially in the case of very small poisoned ratios. 3) Since our screening strategy does not need to be combined with the training process, which greatly saves computational resources and time costs, it can be used as a pre-processing stage of the training set in conjunction with a variety of other attack methods to improve the efficiency of the attack. To support point 3), in addition to BadNets [10], we also combined our HFE-based strategy with the latest Blended Attack [12], InputAware (IW) Attack [31], SIG Attack [26], Label-Consistant (LC) Attack [44], SSBA [28], FTrojan [37] and Low Frequency (LF) attack [21] on CIFAR-10 respectively. The effectiveness of our proposed HFE-based method is verified by the experimental results of multiple attack methods, which can be seen in Figure 6. As shown in the figure, whether the attack pattern is poison-label or clean-label, visible or invisible, poisoning-based or non-poisoning-based, spatial-domain attack or frequency-domain attack, our HFE-based method has greatly helped to improve the attack success rate. In particular, it can be seen that the ASR improvement is more significant when the proportion of sample poisoning is smaller. This is because as the poisoning proportion r increases, the poisoned

TABLE I

THE MEAN ATTACK SUCCESS RATE (%) AND THE OFFSETTING VALUES ON CIFAR-10 COMPARED WITH RS AND FUS

methods \ r		0.001	0.002	0.003	0.004	0.005	0.01	0.015	0.02	0.025	Time
ResNet-18	RS	36.21 ± 0.32	53.28 ± 0.28	61.87 ± 0.39	60.02 ± 0.24	77.82 ± 0.16	87.61 ± 0.38	93.26 ± 0.31	94.13 ± 0.34	95.62 ± 0.19	$< 5s$
	FUS	48.17 ± 0.27	**57.32** ± 0.35	66.76 ± 0.26	66.28 ± 0.49	80.97 ± 0.31	92.13 ± 0.29	96.68 ± 0.22	97.74 ± 0.39	**98.51** ± 0.38	$12h$
	HFE	**49.02** ± 0.34	57.28 ± 0.23	**68.25** ± 0.37	**69.84** ± 0.36	**82.34** ± 0.33	**97.05** ± 0.26	**98.39** ± 0.43	**98.86** ± 0.36	98.37 ± 0.24	$10s$
ResNet-34	RS	54.59 ± 0.29	72.61 ± 0.22	93.14 ± 0.54	95.38 ± 0.46	94.74 ± 0.53	96.23 ± 0.23	97.09 ± 0.35	97.78 ± 0.27	97.92 ± 0.39	$< 5s$
	FUS	69.73 ± 0.32	80.27 ± 0.15	95.49 ± 0.52	95.77 ± 0.22	95.92 ± 0.43	97.85 ± 0.27	**98.42** ± 0.36	98.67 ± 0.29	**98.82** ± 0.31	$12h$
	HFE	**72.86** ± 0.47	**83.96** ± 0.29	**96.17** ± 0.26	**96.28** ± 0.38	**96.27** ± 0.34	**98.24** ± 0.37	98.33 ± 0.26	**98.72** ± 0.13	98.76 ± 0.22	$10s$
VGG-13	RS	53.26 ± 0.23	80.31 ± 0.35	86.82 ± 0.52	88.27 ± 0.19	88.19 ± 0.64	91.72 ± 0.23	92.88 ± 0.31	94.35 ± 0.24	94.42 ± 0.39	$< 5s$
	FUS	62.35 ± 0.47	82.77 ± 0.26	88.26 ± 0.46	**90.76** ± 0.31	90.98 ± 0.13	92.35 ± 0.37	**93.83** ± 0.43	94.96 ± 0.17	96.33 ± 0.22	$12h$
	HFE	**67.28** ± 0.24	**83.69** ± 0.27	**88.44** ± 0.61	90.38 ± 0.25	**92.64** ± 0.31	**93.19** ± 0.29	93.78 ± 0.22	**96.47** ± 0.42	**96.72** ± 0.19	$10s$
VGG-16	RS	62.13 ± 0.42	83.92 ± 0.26	86.97 ± 0.23	88.69 ± 0.18	87.64 ± 0.36	91.33 ± 0.44	92.98 ± 0.52	94.46 ± 0.24	95.17 ± 0.30	$< 5s$
	FUS	73.89 ± 0.33	**86.62** ± 0.34	87.98 ± 0.29	89.15 ± 0.38	89.82 ± 0.27	91.79 ± 0.42	93.44 ± 0.37	94.98 ± 0.26	**97.25** ± 0.19	$12h$
	HFE	**77.26** ± 0.35	85.48 ± 0.27	**89.37** ± 0.35	**90.04** ± 0.22	**91.77** ± 0.31	**93.82** ± 0.37	**93.78** ± 0.41	**96.72** ± 0.36	97.13 ± 0.27	$10s$

TABLE II

TEST MEAN ASR (%) UNDER ALL-TO-ALL SETTING

r	0.005	0.01	0.015	0.02	0.025
RS	13.36	20.23	15.13	53.02	66.27
HFE	**17.61**	**21.2**	**40.81**	**65.92**	**69.24**

TABLE III

THE MEAN ASR (%) ON CIFAR-10 WITH DIFFERENT TARGET LABELS

target label	airplane	dog	horse	ship	truck
RS	87.66	88.14	87.15	88.93	89.31
HFE	**97.08**	**94.37**	**93.72**	**95.26**	**93.80**

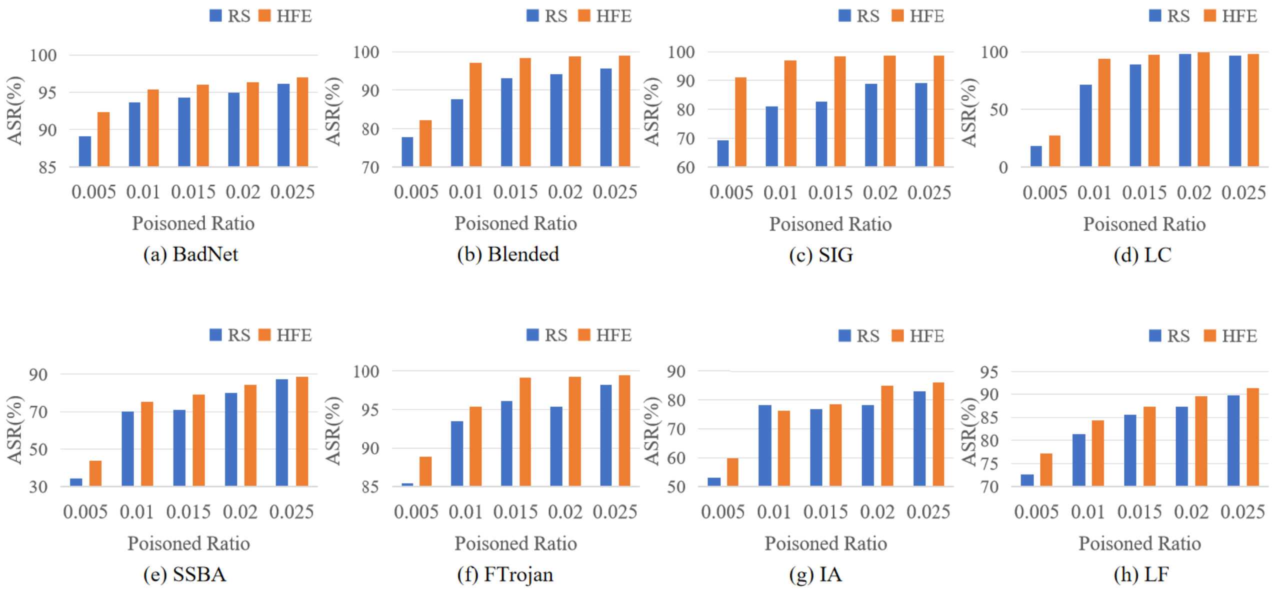
sample set will contain more and more samples with lower contribution, which makes the decision boundary gradually move away from the optimal solution.

2) Different Target-Class Setting: By the way, we conducted the above series of experiments under *all-to-one* setting, where the target category is one of the specified categories. We also supplement the experimental results under the *all-to-all* setting in Table II and some other specified target labels under the all-to-one setting in Table III on CIFAR-10. Even under the tougher all-to-all setting, our HFE-based screening strategy can also improve the attack success rate to a certain extent. This once again confirms that our proposed strategy can be combined with many types of backdoor attack methods to improve the attack effectiveness, especially in the case of a large collective volume of training data, which gives our strategy an even greater advantage in terms of time and computational costs.

3) Defensive Performance: To demonstrate the efficiency of the screened high-contribution poisoned samples with our HFE-based data-screening method, we also conducted defense performance testing using a number of traditional backdoor defense techniques, including Fine-Tuning (FT), Fine-Pruning (FP) [45], Neural Attention Distillation (NAD) [46], Activation Clustering (AC) [47], Anti-Backdoor Learning (ABL) [48] and Spectral Signatures (SP) [49]. Fine-tuning (FT) is a common defense that involves a small amount of local retraining on a clean training set to perform micro-tuning of model parameters,thus providing some degree of protection. Since backdoor attacks utilize spare capacity in neural networks, Pruning [50], again as a common defense, reduces the size of the backdoored network by pruning neurons that would lie dormant in the face of benign inputs, ultimately rendering the backdoor behavior ineffective. FP [45] combines the advantages of Fine-Tuning and Pruning, which can more effectively weaken or eliminate backdoor attacks. They first prune the DNN, and then fine-tune the pruned network. NAD [46] uses a teacher model to guide a backdoor student model in fine-tuning a small clean subset so that the middle layer attention of the student model is aligned with that of the teacher network, which in turn mitigates backdoor attacks. AC [47] detects the backdoor samples inserted into the DNN and then analyzes the neural network activation state of the training samples to determine whether they are poisoned or not. ABL [48] introduces a two-stage gradient ascent mechanism designed to train a clean model on poisoned samples. The first stage it incorporates a local gradient ascent mechanism for isolating a small fraction of poisoned samples from the training data. The second stage it utilizes the global gradient ascent mechanism to break the strong correlation between the poisoned samples and the attack target class. And SP [49] shows that backdoor attacks tend to leave behind a detectable trace, also called a spectral signature, in the spectrum of the covariance of a feature representation learned by the neural network. And they demonstrate that one can use this signature to identify and remove the poisoned inputs. It is worth mentioning that we are grateful for the BackdoorBench[1] of Wu et al. [51]. We used their open code resources to conduct most of the defense tests in the experimental section, which saved a lot of time for the validation of our proposed method.

In this set of experiments, we use RS and our HFE-based method to screen out a specific percentage of poisoned samples, respectively, and then use these poisoned samples to perform BadNets attack [10]. The dataset we used is CIFAR-10, and the model structure we used is ResNet18. The specific experimental results are shown in Figure 7. For this set of experiments, at the same poisoning ratio r, a higher ASR means that the selected poisoned samples are more resistant to defense. From the analysis of these experimental results, our method screens out highly poisoned samples that are critical to the formation of decision boundaries and greatly improve the resistance of the samples to defense.

[1] https://github.com/SCLBD/BackdoorBench



(a) BadNet (b) Blended (c) SIG (d) LC

(e) SSBA (f) FTrojan (g) IA (h) LF

Fig. 6. The mean ASR (%) using Random Selection (RS) strategy and our HFE-based strategy respectively under different attack methods. The dataset we used is CIFAR-10, and the model is ResNet-18. The horizontal axis is the proportion of poisoned samples screened, and the vertical axis is the corresponding attack success rate.

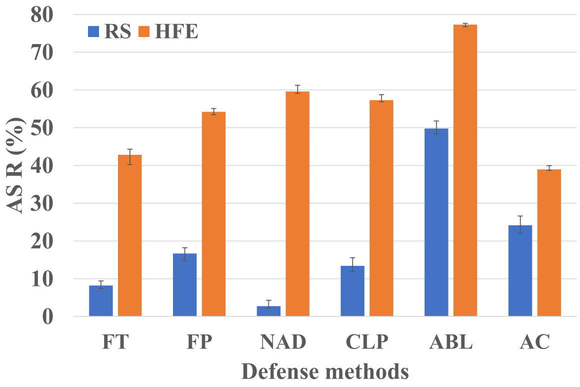


Fig. 7. The mean attack success rate (%) on CIFAR-10 with different defense methods. The attack method we used is BadNets [10]. The model structure is RseNet18. And the R_{hp} we applied in HFE defaults to 10.

TABLE IV
THE ANOMALY INDEX USING NEURAL CLEANSE

r	0.005	0.01	0.015	0.02	0.025	0.03
RS	3.806	3.722	3.853	4.006	6.044	7.434
HFE	**2.303**	**2.447**	**2.628**	**3.908**	**2.860**	**5.474**

TABLE V
THE DETECTION ACCURACY (%) USING AEVA

r	0.005	0.01	0.015	0.02	0.025
RS	**32.3**	39.6	**42.4**	43.7	49.5
HFE	31.2	**40.7**	41.9	**44.2**	**49.6**

In order to explore whether our HFE-based data-screening strategy would make these efficient samples outliers and thus easily detectable by some detection algorithms, we used two classic backdoor detection algorithms, Neural Cleanse (NC) [52] and AEVA [53], to investigate and found that this concern is redundant. As shown in Table IV below, the metric Anomaly Index (A_idx) is used to determine whether the model is backdoored. When $A\_idx < 2$, it is a clean model, otherwise, it is poisoned. And the larger A_idx the higher the poisoning probability. From the results, we can find: 1) overall the smaller the poisoning samples, the lower the anomaly index of the model, which implies a reduction in the stealthiness of the attacks; 2) Compared to RS, the samples selected by our HFE will not be more easily detected. Please note that we are not simply selecting samples with higher frequency energy, but are concerned with the **changes** in HFE after adding triggers. Since the changes are also located within the normal frequency domain distribution, the screened samples by our method will not become outliers. The results of detection accuracy using AEVA methods are shown in Table V, which shows that our screening strategy has little effect on detection accuracy and does not make the samples more likely to be detected by the detection algorithms.

C. Ablation Study

1) The Impact of the High-Pass Filter Radius R_{hp}: As the size of the transformed amplitude spectrum is the same as the original, both are 32×32. And the frequency domain can be divided into three frequency bands: low frequency, medium frequency, and high frequency. In order to investigate the most suitable high-pass filtering radius R_{hp}, we carry out extensive experiments for this parameter. We explore the effect of R_{hp} on the final attack success rate on two datasets, CIFAR10 and CIFAR100, respectively. And the results are shown in Table VI and Table VII. The target class in both sets of experiments is 0. And both of them use ResNet-18 as their training model architecture. As can be seen from the results, our method achieves higher performance compared to both

TABLE VI

The Attack Success Rate (%) on CIFAR-10 With Different R_{hp}. R_{hp} Is the Radius of Our High-Pass Filter Used in the HFE Extraction Process. r Is the Poisoned Ratio. RS Means Random Selection Strategy of Poisoned Samples

R_{hp} \ r	0.005	0.01	0.015	0.02	0.025	0.03
RS	77.8	87.6	93.2	94.1	94.9	96.3
HFE (R_{hp}=8)	80.8	94.3	97.5	98.5	98.7	98.9
HFE (R_{hp}=10)	81.2	94.1	97.6	98.6	98.6	**99.1**
HFE (R_{hp}=12)	**82.3**	**97.0**	**98.3**	**98.8**	**98.9**	**99.1**
HFE (R_{hp}=14)	80.9	96.0	97.1	98.4	**98.9**	99.0

TABLE VII

The Attack Success Rate (%) on CIFAR-100 With Different R_{hp}. R_{hp} Is the Radius of Our High-Pass Filter Used in the HFE Extraction Process. r Is the Poisoned Ratio. RS Means Random Selection Strategy of Poisoned Samples

methods \ r	0.005	0.01	0.015	0.02	0.025	0.03
RS	63.2	78.5	86.7	90.5	92.2	94.1
HFE (R_{hp}=8)	66.9	82.7	90.9	92.7	94.3	95.6
HFE (R_{hp}=10)	68.1	82.9	**91.2**	**92.9**	**94.5**	**96.4**
HFE (R_{hp}=12)	**70.1**	**84.3**	90.3	92.2	94.3	95.8
HFE (R_{hp}=14)	69.4	83.7	90.3	92.6	94.4	96.1

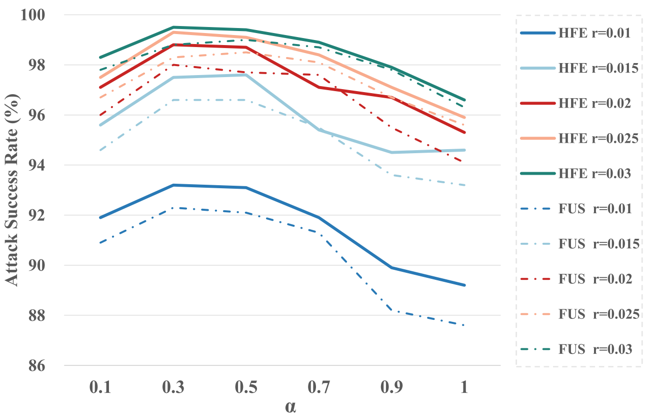


Fig. 8. The mean ASR (%) on CIFAR-10 with different α compared with FUS. As shown in the legend, the solid and dashed lines indicate our HFE and FUS methods respectively, and different colors indicate different poisoning ratios.

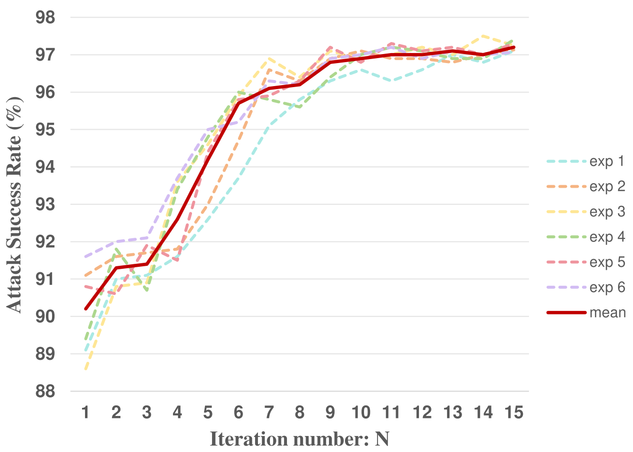


Fig. 9. The attack success rate (%) on CIFAR-10 with different iteration numbers N. And the solid line in the figure represents the mean ASR values of the six-group experiments.

RS and FUS [13] based on forgotten events. The best overall attack success rate was achieved when R_{hp} is around $10 \sim 12$.

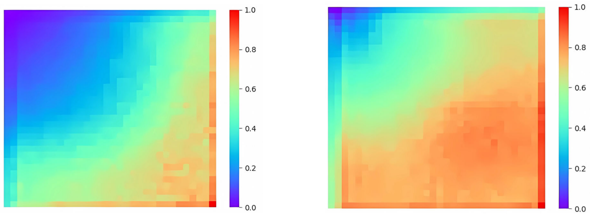
We believe that when R_{hp} is too large or too small, it will destroy the integrity of the high-frequency information, which in turn affects the sensitivity of the network to the high-frequency information of the image. When R_{hp} is too small, some low-frequency information is mixed in the obtained HFE. When R_{hp} is too large, the obtained high-frequency information is again not complete. Therefore, the efficiency of the method can be maximized only when R_{hp} is moderate. Generally speaking, the high-pass filter's radius R_{hp} is about $1/3$ of the original image size is more appropriate.

2) The Size of $\mathcal{D}'_p$ Controlled by α: To investigate the effect of the size of the alternative sample pool $|\mathcal{D}'_p| = \alpha \cdot |\mathcal{U}|$ on the validity of the final sample screening, we conducted this set of experiments on CIFAR10 for different α. We compared our HFE-based strategy with FUS [13] under different settings of poisoned ratio $r : 0.01, 0.015, 0.02, 0.025, 0.03$. The model we used is ResNet-18. R_{hp} defaults to 12. The results are shown in Figure 8. We can find that α is either too small or too large resulting in a performance decrease, with the former leading to a slower update of the sample pool and the latter to an algorithmic convergence failure. As seen in the results, the global screening strategy works even better when α is set at about 0.3 numerically.

3) The Iteration Number N: In the process of updating the data screening strategy, it is important to consider the iteration number N as it directly impacts the globalization of the samples. A smaller iteration number results in poorer globalization of the samples. To investigate this, we conducted six sets of experiments and calculated the average values. The results, as depicted in Figure 9, indicate that stronger globalization of the filtered samples leads to a higher attack success rate. Additionally, as the iteration number N increases, the globalization reaches a saturation point, resulting in the attack success rate gradually converging and stabilizing. This experimental exploration highlights the significance of iteration number N in achieving effective data screening.

D. Attribution Studies

In this part, we analyze the experimental results from both quantitative and qualitative perspectives. To demonstrate the effectiveness of our choice of HFE rather than low-frequency energy (LFE) during the data screening phase, we compared the effect of low-pass filtering and high-pass filtering on the final attack success rate. In this set of experiments, we use CIFAR-10 and choose ResNet-18 as our model architecture. The backdoor attack method is also Blended Attack [12] and the target label is 0. During the data screening phase, R_{hp} defaults to 12, and α defaults to 0.3. We conducted this set of experiments for different poisoning ratios r : $0.005, 0.01, 0.015, 0.02, 0.025, 0.03$. And the detailed results are shown in Figure 11. From the results in Figure 11, we can find that there is not much difference between the attack



(a) CIFAR-10 (b) CIFAR-100

Fig. 10. The heat map of the gradient back-propagation in the frequency domain performed BadNets [10]. The upper left corner represents the lowest frequency and the lower right corner represents the highest frequency.

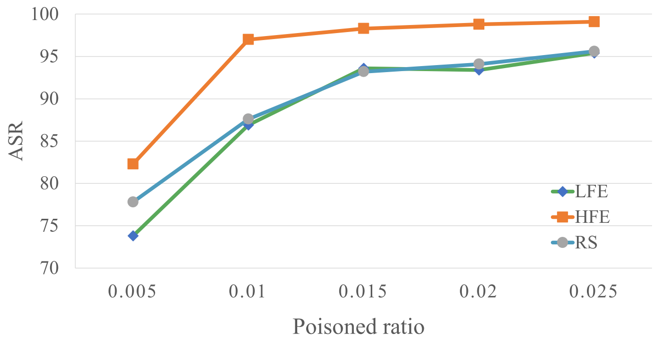


Fig. 11. The mean ASR (%) on CIFAR-10 with different frequency energy statistics approaches during the data screening. LFE/HFE means that we use low-/high-frequency energy to rank the samples. And RS means random selection strategy of poisoned samples.

success rate (ASR) of using LFE and RS strategies, and even the ASR corresponding to LFE decreases when r is 0.01, 0.015 and 0.025. As for the reason for this phenomenon in the qualitative analysis above, we believe that the samples filtered by LFE have less effective HFE than the RS, which in turn has less impact on the final decision boundary of the classification network. In contrast, the samples screened with HFE are more efficient in backdoor attacks. During the gradient back-propagation, we employ SSA (Spectrum Simulation Attack) [54] to visualize the network's heat map in the frequency domain, as shown in Figure 10. We perform BadNets attacks on CIFAR-10 and CIFAR-100, respectively. The model structure we used is ResNet18. In this figure, the frequency component is lower in the upper-left part and higher in the upper-right part. And the warmer the color means that the model pays more attention to that part of the region. This is also strong proof of the intuition of our proposed HFE-based method.

V. Limitations

There are two primary limitations in our proposed methodology as well as in the experimental validation. Firstly, we have conducted performance testing using images of sizes 32×32 and 224×224, without validating the performance on higher resolution, larger volume, and more complex datasets. In future studies, we will investigate the effectiveness of diverse types of datasets to enhance our findings. Secondly, this paper solely focuses on the image classification task. In the future, we aim to explore the potential utilization of frequency domain information in more intricate deep vision tasks.

VI. Conclusion

In this paper, we propose a minimalist data screening strategy utilizing the change of HFE after poisoning to screen out the high-contribution poisoned samples. Compared with the previous method, our proposed method not only greatly improves the time efficiency, but also increases the ASR of both poison-label and clean-label backdoor attacks. And the test results of multiple defense methods show that the poisoned samples screened out by our HFE-based method have higher resistance to defense.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:195908774

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[3] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1–9. [Online]. Available: https://api.semanticscholar.org/CorpusID:206592484

[4] C. Szegedy, A. Toshev, and D. Erhan, "Object detection," in *Proc. Neural Inf. Process. Syst.*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:8827762

[5] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:152282225

[6] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:49862415

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440. [Online]. Available: https://api.semanticscholar.org/CorpusID:1629541

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 731–737.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[10] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:131777414

[11] Y. Liu et al., "Trojaning attack on neural networks," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:31806516

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

[13] P. Xia, Z. Li, W. Zhang, and B. Li, "Data-efficient backdoor attacks," 2022, *arXiv:2204.12281*.

[14] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, "Training behavior of deep neural network in frequency domain," in *Proc. Int. Conf. Neural Inf. Process.*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49562099

[15] Z.-Q. John Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

[16] T. Luo, Z. Ma, Z.-Q. J. Xu, and Y. Zhang, "Theory of the frequency principle for general deep neural networks," 2019, *arXiv:1906.09235*.

[17] Y. Zhang, Z.-Q. J. Xu, T. Luo, and Z. Ma, "Explicitizing an implicit bias of the frequency principle in two-layer neural networks," 2019, *arXiv:1905.10264*.

[18] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A Fourier perspective on model robustness in computer vision," in *Proc. Neural Inf. Process. Syst.*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195317007

[19] H. Wang, X. Wu, Z. Huang, and E. P. Xing, “High-frequency component helps explain the generalization of convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8681–8691.

[20] N. Rahaman et al., “On the spectral bias of neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53012119

[21] Y. Zeng, W. Park, Z. M. Mao, and R. Jia, “Rethinking the backdoor attacks’ triggers: A frequency perspective,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 16453–16461. [Online]. Available: https://api.semanticscholar.org/CorpusID:233182042

[22] Y. Liu, X. Ma, J. Bailey, and F. Lu, “Reflection backdoor: A natural backdoor attack on deep neural networks,” 2020, *arXiv:2007.02343*.

[23] S. Li, M. Xue, B. Zhao, H. Zhu, and X. Zhang, “Invisible backdoor attacks on deep neural networks via steganography and regularization,” *IEEE Trans. Dependable Secure Comput.*, vol. 18, pp. 2088–2105, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:220633516

[24] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, “Clean-label backdoor attacks on video recognition models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14431–14440. [Online]. Available: https://api.semanticscholar.org/CorpusID:212628208

[25] A. Shafahi et al., “Poison frogs! Targeted clean-label poisoning attacks on neural networks,” in *Proc. Neural Inf. Process. Syst.*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4626477

[26] M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in CNNs by training set corruption without label poisoning,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 101–105. [Online]. Available: https://api.semanticscholar.org/CorpusID:67855469

[27] E. Sarkar, H. Benkraouda, G. Krishnan, H. Gamil, and M. Maniatakos, “FaceHack: Attacking facial recognition systems using malicious facial characteristics,” *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 3, pp. 361–372, Jul. 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:244867412

[28] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, “Invisible backdoor attack with sample-specific triggers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16443–16452. [Online]. Available: https://api.semanticscholar.org/CorpusID:237054216

[29] H. Souri, L. Fowl, R. Chellappa, M. Goldblum, and T. Goldstein, “Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch,” 2021, *arXiv:2106.08970*.

[30] E. Bagdasaryan and V. Shmatikov, “Blind backdoors in deep learning models,” in *Proc. USENIX Secur. Symp.*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218571440

[31] A. Nguyen and A. Tran, “Input-aware dynamic backdoor attack,” 2020, *arXiv:2010.08138*.

[32] K. D. Doan and Y. Lao, “Backdoor attack with imperceptible input and latent modification,” in *Proc. Neural Inf. Process. Syst.*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245011163

[33] K. Doan, Y. Lao, W. Zhao, and P. Li, “LIRA: Learnable, imperceptible and robust backdoor attacks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11946–11956. [Online]. Available: https://api.semanticscholar.org/CorpusID:244397177

[34] T. A. Nguyen and A. T. Tran, “WaNet—Imperceptible warping-based backdoor attack,” in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=eEn8KTtJOx

[35] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu, and D. Miller, “Backdoor embedding in convolutional neural network models via invisible perturbation,” in *Proc. 10th ACM Conf. Data Appl. Secur. Privacy*, Mar. 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:52138086

[36] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 86–94. [Online]. Available: https://api.semanticscholar.org/CorpusID:11558223

[37] T. Wang, Y. Yao, F. Xu, S. An, H. Tong, and T. Wang, “An invisible black-box backdoor attack through frequency domain,” in *Proc. Eur. Conf. Comput. Vis.*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253448651

[38] H. A. A. K. Hammoud and B. Ghanem, “Check your other door! Creating backdoor attacks in the frequency domain,” 2109, *arXiv:2109.05507*.

[39] T. Wang, Y. Yao, F. Xu, S. An, H. Tong, and T. Wang, “Backdoor attack through frequency domain,” 2021, *arXiv:2111.10991*.

[40] H. A. Al Kader Hammoud, A. Bibi, P. H. S. Torr, and B. Ghanem, “Don’t FREAK out: A frequency-inspired approach to detecting backdoor poisoned samples in DNNs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2338–2345. [Online]. Available: https://api.semanticscholar.org/CorpusID:257687422

[41] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, “Backdoor learning: A survey,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:220633116

[42] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Tech. Rep., 2009.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255. [Online]. Available: https://api.semanticscholar.org/CorpusID:57246310

[44] A. Turner, D. Tsipras, and A. Madry, “Label-consistent backdoor attacks,” 2019, *arXiv:1912.02771*.

[45] K. Liu, B. Dolan-Gavitt, and S. Garg, “Fine-pruning: Defending against backdooring attacks on deep neural networks,” 2018, *arXiv:1805.12185*.

[46] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, “Neural attention distillation: Erasing backdoor triggers from deep neural networks,” 2021, *arXiv:2101.05930*.

[47] B. Chen et al., “Detecting backdoor attacks on deep neural networks by activation clustering,” 2018, *arXiv:1811.03728*.

[48] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, “Anti-backdoor learning: Training clean models on poisoned data,” in *Proc. Neural Inf. Process. Syst.*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239616453

[49] B. Tran, J. Li, and A. Madry, “Spectral signatures in backdoor attacks,” in *Proc. Neural Inf. Process. Syst.*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53298804

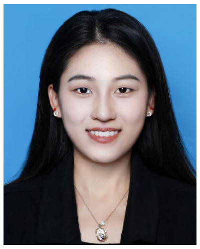
[50] S. Han, J. Pool, J. Tran, and W. J. Dally, “Learning both weights and connections for efficient neural network,” in *Proc. Neural Inf. Process. Syst.*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:2238772

[51] B. Wu et al., “BackdoorBench: A comprehensive benchmark of backdoor learning,” 2022, *arXiv:2206.12654*.

[52] B. Wang et al., “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 707–723. [Online]. Available: https://api.semanticscholar.org/CorpusID:67846878

[53] J. Guo, A. Li, and C. Liu, “AEVA: Black-box backdoor detection using adversarial extreme value analysis,” 2021, *arXiv:2110.14880*.

[54] Y. Long et al., “Frequency domain model augmentation for adversarial attack,” 2022, *arXiv:2207.05382*.



Yuan Xun is currently pursuing the joint Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing. Her research interests include computer vision, deep learning, and adversarial machine learning.

Xiaojun Jia received the joint Ph.D. degree from the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing. He is currently a Research Fellow with the Cyber Security Research Centre @ NTU, Nanyang Technological University, Singapore. His research interests include computer vision, deep learning, and adversarial machine learning.

Jindong Gu (Member, IEEE) received the Ph.D. degree from the University of Munich, advised by Prof. Volker Tresp. He has worked/interned with Tencent AI Lab, Microsoft Research, and Google Research. He is currently a Post-Doctoral Researcher with the University of Oxford, working with Prof. Philip Torr. His long-term research goal is to build responsible general intelligence systems.

Xinwei Liu is currently pursuing the joint Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing. His research interests include computer vision, deep learning, and adversarial machine learning.

Qing Guo (Member, IEEE) received the Ph.D. degree in computer application technology from the School of Computer Science and Technology, Tianjin University, China. He is currently a Senior Research Scientist and the Principal Investigator (PI) of the Centre for Frontier AI Research (CFAR), A*STAR, Singapore. He is also an Adjunct Assistant Professor with the National University of Singapore (NUS). Before that, he was a Wallenberg-NTU Presidential Post-Doctoral Fellow with Nanyang Technological University, Singapore. His research interests include computer vision, AI security, and image processing. He is a Senior PC Member of AAAI.

Xiaochun Cao (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. After graduation, he spent about three years with ObjectVideo Inc., as a Research Scientist. He is currently with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China. He has authored and coauthored more than 100 journal and conference papers. He is a fellow of IET. His dissertation was nominated for the University of Central Florida's university-level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award from the International Conference on Pattern Recognition. He is on the Editorial Board of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.