Article | Received 4 May 2026; Revised 3 June 2026; Accepted 17 June 2026; Published 24 July 2026
https://doi.org/10.55092/aic20260011

FBFS: fog backdoor attack based on feature similarity

Bin Zhou[1], Wenchao Zhang[1], Hongao Yang[1], Guodong Ye[2] and Xingxing Jia[1,*]

[1] School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, China
[2] School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China

* Corresponding author; Email: jiaxx@lzu.edu.cn.

Highlights:

- Introduce FBFS: a fog-based backdoor attack method with high stealthiness.
- Use a physical fog model as a visually plausible trigger in images.
- Apply feature similarity regularization to align poisoned and clean samples.
- Achieve high attack success rate while maintaining clean sample accuracy.

Abstract: Most existing backdoor attacks embed triggers in images in ways that are either conspicuous to human observers or easily detected by feature-space defenses, thereby sacrificing stealthiness or robustness. To address these issues, we propose a fog backdoor attack based on feature similarity (FBFS), which enhances the visual concealment of backdoor triggers as well as the feature-space homogeneity between poisoned and clean samples. Specifically, FBFS employs a standard optical model to simulate natural fog as a visually plausible trigger and injects it into image samples. Additionally, a feature similarity penalty term is incorporated into the loss function to enforce consistency in the feature representations of poisoned and clean samples, thereby evading defenses that rely on latent separability. Experiments conducted on Canadian Institute for Advanced Research 10-class dataset (CIFAR-10), German Traffic Sign Recognition Benchmark (GTSRB), and a subset of ImageNet demonstrate that, under a 10% poisoning rate, FBFS achieves over 90% attack success rate while maintaining clean sample accuracy above 85%. Moreover, detection rates under representative feature-space defense methods, including activation clustering and spectral signature analysis, remain below 40%, demonstrating that the proposed method effectively balances attack performance and resistance to detection, exhibiting both stealth and robustness.

Keywords: backdoor attack; feature similarity; visual stealth; feature space; latent separability; model robustness

1. Introduction

In recent years, the rapid development of the internet industry and artificial intelligence technology has led to significant advancements in deep neural networks for tasks such as image recognition and natural language processing. However, the training process of models increasingly relies on third-party

platforms and cloud services, a trend that has raised widespread concerns regarding model security. Attackers can manipulate training samples or the training pipeline to implant backdoor into models, thereby maintaining normal performance while causing the model to exhibit predefined behaviors under specific triggers, posing serious security risks.

In 2017, Gu *et al.* [1] first introduced the concept of backdoor attacks and designed the classic BadNets attack scheme. This method involves embedding an attacker-prescribed visual trigger into a portion of the training samples while simultaneously altering their labels to the target class, thereby achieving malicious control over the model. As shown in Figure 1, the model maintains its original performance on normal inputs but outputs the attacker-specified class when the input contains the specific trigger pattern. The BadNets attack not only achieves a high attack success rate but also barely affects the prediction accuracy on clean data, thus exhibiting strong stealthiness. Since model users typically only possess clean datasets and cannot trigger the backdoor behavior, it becomes difficult for them to detect the hidden security risks within the model. Backdoor attacks have also been applied to various scenarios, such as federated learning [2–4], Graph Neural Networks (GNNs) [5,6], and malicious traffic evasion [7]. These characteristics of "controllable input conditions" and broad application scenarios make backdoor attacks one of the most threatening security vulnerabilities in deep learning systems.

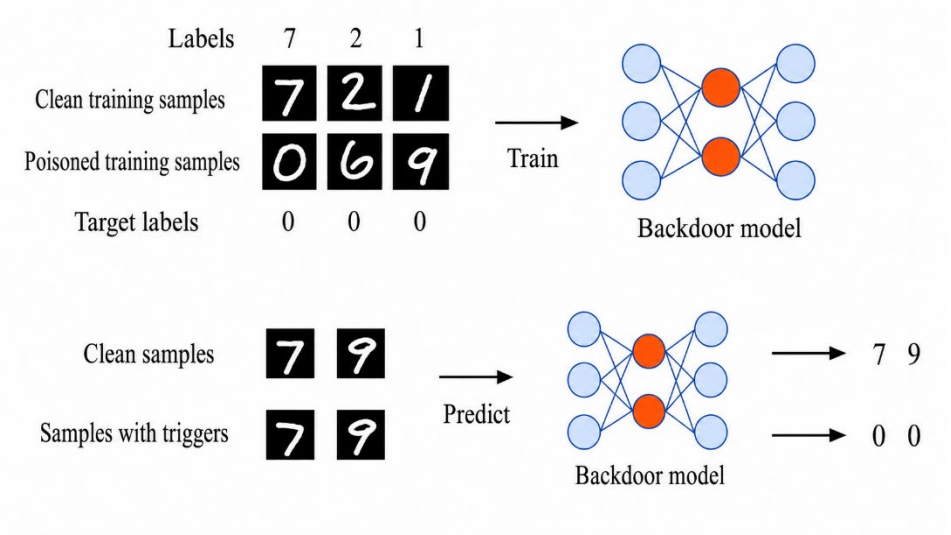

Figure 1. The implementation principle of backdoor attacks.

Recent studies indicate that backdoor threats are extending beyond discriminative classifiers to advanced generative models. Han *et al.* proposed UIBDiffusion, a universal imperceptible backdoor attack against diffusion models, demonstrating that minor adversarial perturbations can remain highly stealthy while maintaining attack effectiveness. This trend suggests that effective backdoor triggers should simultaneously ensure imperceptibility, semantic plausibility, and robustness under practical transformations, which motivates the fog-based trigger design in fog-based backdoor attack with feature similarity (FBFS) [8].

However, existing research has found that most backdoor attacks exhibit the potential separability between poisoned samples and clean samples in the feature space, which becomes the key basis for feature-space detection methods (such as activation clustering, spectral signatures, *etc.*) to identify backdoored models. To counteract this issue, Qi *et al.* [9] proposed the Adapted Patched method, which

reduces potential separability through asymmetric trigger design. However, this method involves a complex training process, making it difficult to deploy.

To address this, this paper proposes a fog-based backdoor attack method FBFS that leverages dual stealthiness design in both visual and feature spaces to enhance the concealment and robustness of backdoor attacks. The main contributions of this paper are as follows:

(1) We propose a backdoor attack method, FBFS, that integrates physical fog modeling and feature similarity constraints. This method employs a standard optical model to generate natural fog, which is embedded into training samples as the image trigger. This approach not only exhibits strong visual stealthiness but also demonstrates high plausibility due to its consistency with real-world scenes, thereby significantly reducing the probability that a human observer will detect it.

(2) We propose a feature-space mixing strategy based on feature similarity to evade existing detection mechanisms. A penalty term is introduced into the loss function to constrain the distribution of toxic samples and clean samples, ensuring they are as uniform as possible in the feature space. This avoids easy detection by defense methods specifically designed for the feature space, thereby enhancing the anti-detection capability and robustness of the backdoored model.

(3) We conduct systematic empirical validation on multiple datasets. Comparative experiments with typical backdoor methods such as BadNets and Adapted Patched on CIFAR-10, GTSRB, and a subset of ImageNet show that FBFS can effectively carry out backdoor attacks while maintaining the model's normal performance. It also enhances the model's robustness and effectively evades defense methods based on feature-space design.

2. Related work

2.1. Evolution of backdoor attack paradigms and trigger mechanisms

Backdoor attacks represent a class of security threats in deep learning systems that involve injecting malicious behaviors during the model training phase. These attacks typically modify a subset of training samples by altering their inputs and corresponding labels, causing the model to output attacker-specified results when presented with inputs containing specific triggers. Such attacks are characterized by high stealthiness and deployment flexibility and have been applied across various domains, including image, text, and speech processing [10–14].

Early backdoor attack schemes, exemplified by BadNets [1], embedded fixed patterns (e.g., white squares) into a portion of training samples and modified their labels to implant backdoor behaviors into the model. These methods are simple in structure and exhibit stable attack performance, but the triggers are easily detectable by the human eye or detection models, resulting in poor stealthiness. In recent years, with advancements in detection techniques, researchers have gradually shifted from "fixed pattern triggers" [1] to designs with greater stealth and environmental adaptability, proposing ideas such as invisible triggers [10], natural triggers [11], and input-aware triggers. For instance, Refool [11] leveraged natural optical simulations like reflections to embed backdoor patterns while preserving image semantic consistency, significantly enhancing visual stealthiness. Furthermore, for multi-target attacks, we utilize Discrete Cosine Transform (DCT) steganography to construct the trigger in the frequency domain and embed it into different channels of the image [12]. This approach achieves strong results in both

imperceptible N-to-N (multi-target) backdoor attacks and imperceptible N-to-One (multi-trigger) backdoor attacks.

Regarding label manipulation strategies, beyond the classic targeted attacks with label modification, new variants such as clean-label attacks [13,14] and sample-specific attacks [15,16] have emerged, further increasing the unpredictability and detection difficulty of attacks. Recent studies have further extended backdoor attacks from conventional CNN classifiers to semi-supervised learning, physical-world deployment, vision-language models, transformer architectures, and non-Euclidean data. Gong *et al.* [17] proposed Megatron, an evasive clean-label backdoor attack against vision transformers, which leverages latent alignment and attention diffusion to improve both attack effectiveness and stealthiness. Dai *et al.* [18] investigated unnoticeable backdoor attacks on graph neural networks, showing that poisoned-node selection and adaptive trigger generation can enhance attack efficacy while maintaining stealthiness under limited attack budgets. Feng *et al.* [19] showed that semi-supervised models trained on unlabeled web images can still be compromised through gradient-matched poisoning. Guo *et al.* [20] proposed a physically invisible backdoor based on camera imaging, where the trigger is associated with device-specific camera fingerprints rather than explicit pixel perturbations. Li *et al.* [21] further demonstrated object-oriented backdoor attacks against image captioning, highlighting that backdoor threats are rapidly extending to vision-language systems and more realistic deployment settings. To mitigate the potential separability of backdoor attacks, Qi *et al.* [9] proposed the Adapted Patched attack, which employs an asymmetric trigger strategy to reduce the potential separability between poisoned and clean samples in the feature space, thereby attempting to evade clustering-based detection methods. This research highlights a new challenge in backdoor attacks: how to enhance attack efficiency while effectively evading feature-space defense mechanisms.

2.2. Feature space detection mechanisms and defense methods

Mainstream backdoor defense methods often rely on the anomalies that models exhibit in the feature space, particularly the latent separability between poisoned and clean samples in deep representations. This characteristic has become a critical theoretical foundation for constructing detection algorithms.

Wang *et al.* [22] proposed a classical defense method called Neural Cleanse, which reverse engineers potential triggers in backdoored models to determine the presence of backdoors. They further introduced strategies such as input filtering, pruning backdoor related neurons, and model retraining to prevent backdoor activation. Chen *et al.* [23] developed activation clustering to detect backdoor attacks. They observed that the model's classification mechanisms for clean and poisoned samples differ, resulting in distinct neuron activation patterns. By reducing the dimensionality of activation values for samples of each category, clustering them, and conducting detailed analysis, backdoored models and their target categories can be identified. Tran *et al.* [24] argued that understanding Spectral Signatures is essential for designing machine learning systems resilient to backdoor attacks. Their method leverages the anomalous distribution of poisoned samples in the spectral dimension of features, making it applicable to various model architectures.

These methods share a common premise: the feature distribution of poisoned samples structurally diverges from that of clean samples. Consequently, if attackers actively guide the feature distribution of poisoned samples to align with clean samples during training, such detection mechanisms can be effectively bypassed. Qi *et al.* [9] proposed Adapted Patched preliminarily validated this strategy's

feasibility. However, its high training complexity and strong dependency on specific strategies make it difficult to generalize across multi-task scenarios. More recent defenses have begun to target adaptive attacks that reduce feature-space separability. For example, Guo *et al.* proposed LFPD, which leverages local feature consistency for backdoor detection. This line of work suggests that effective attacks should jointly consider global feature alignment and local visual coherence. Therefore, FBFS adopts a physically plausible fog trigger and feature similarity regularization to mitigate both visual anomalies and latent-space separability [25].

In summary, enhancing the homogeneity in the feature space and rationality in the visual space [5,11–13] has become a key design direction for constructing novel and stealthy backdoor attack methods.

2.3. Positioning and motivation for FBFS

Although existing research has made progress in improving the stealthiness of backdoor attacks, current mainstream methods still face a trade-off between visual camouflage and feature-space robustness. For instance, Adapted Patched suppresses the separability between poisoned and clean samples in the feature space through an asymmetric trigger strategy, thereby evading clustering-based detection methods. However, this approach relies on complex trigger configurations and grouping strategies, resulting in high training costs and poor scalability. Another category of methods, such as Refool [11], which utilizes natural triggers, achieves superior visual stealthiness but fails to optimize the feature-space distribution. As a result, they remain vulnerable to detection mechanisms based on spectral signatures.

In light of these limitations, there is a pressing need for a unified design framework that ensures both trigger naturalness and feature-space blending.

The FBFS method proposed in this paper addresses these challenges by constructing a backdoor attack framework that combines visual plausibility with feature-space stealth. This method employs a standard physical optics model to generate natural fog as a trigger, achieving visual camouflage while preserving image semantics. Additionally, a feature similarity constraint term is incorporated into the loss function to actively guide poisoned samples to aggregate toward clean samples in the feature space, effectively evading detection methods based on distribution heterogeneity. Compared to existing work, FBFS achieves a better balance in trigger construction, training cost, and resistance to detection, demonstrating strong practicality and robustness.

3. Fog-based backdoor attack with feature similarity

To achieve dual stealthiness in both visual and feature spaces, the FBFS method focuses on two key aspects: applying fog using a standard optical model, and designing a penalty term to enhance the distribution homogeneity between backdoored and clean samples in the feature space. This section elaborates on these two components, explaining the underlying principles and implementation step by step.

3.1. Physical fog modeling and trigger generation mechanism

In backdoor attacks, the visual stealthiness of the trigger is crucial for both practical applicability and evasion of detection. Unlike previous approaches that rely on explicit patterns or artificial perturbations, this paper introduces a natural fog model based on physical modeling as the image trigger, achieving a balance between high concealment and scene rationality.

Current image fogging methods primarily fall into three categories: atmospheric scattering-based methods using standard optical models [26], deep learning-based fog synthesis methods using Generative Adversarial Networks (GANs) [27], fog synthesis methods incorporating image depth information. Among these, the standard optical model offers clear physical interpretability, simple implementation, and controllable parameters, making it widely applicable in both dehazing and fog synthesis tasks. In contrast, deep learning-based methods, while capable of learning complex mappings, heavily depend on fog-specific datasets and suffer from limited generalizability and controllability. Methods based on depth information yield superior visual effects but requires costly acquisition of real depth information, restricting their application in general datasets.

This paper adopts an image fogging method based on a standard optical model to generate the backdoor trigger. The model is defined as follows:

$$t(x) = e^{-\beta d(x)} \tag{1}$$

$$Pic(x) = J(x)t(x) + \alpha\big(1 - t(x)\big) \tag{2}$$

where x denotes the pixel coordinate, $J(x)$ is the original clear image, α represents the global atmospheric light, β is the scattering coefficient, and $d(x)$ is the estimated depth value at each pixel. The transmission map $t(x)$, defined as characterizes the attenuation of light due to fog, whose value depends on both scene depth and atmospheric conditions. The fog generation mechanism of the standard optical model is illustrated in Figure 2.

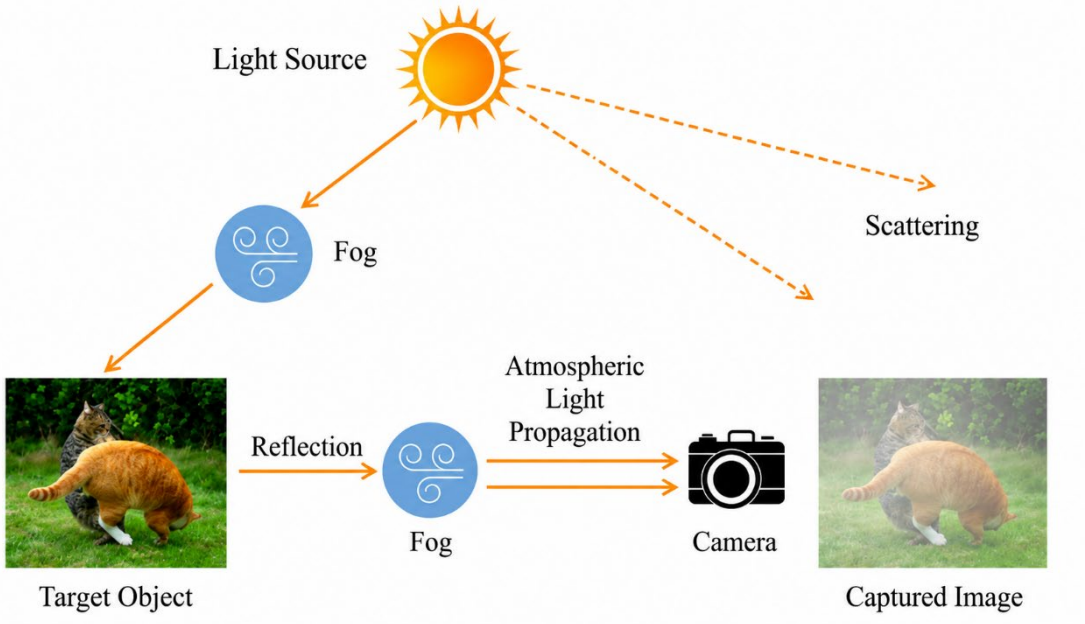

Figure 2. Standard optical model.

This model explains the formation of a foggy image as resulting from two factors: (1) the attenuation of reflected light due to fog, which reduces image brightness, and (2) the scattering effect of atmospheric light, which enhances overall brightness but reduces saturation, thereby producing the characteristic hazy appearance. By setting appropriate scattering coefficients and atmospheric light values, it is possible to simulate natural fog conditions while preserving original image semantics, thereby enabling high-quality local fog generation.

In our experiments, using a subset of ImageNet as an example, simulated fog was injected into random regions of images. Examples of generated samples are shown in Figure 3. It can be observed that such triggers exhibit naturally transitioning fog-like characteristics, highly consistent with real-world visual phenomena. These fog-based triggers demonstrate strong camouflage properties and are difficult to detect with the human eye.

Figure 3. ImageNet's original clean sample (Top) and backdoor sample (bottom).

3.2. Design of feature similarity regularization term and construction of loss function

In backdoor attacks, attackers typically perturb a subset of inputs and manipulate their labels during the training phase, causing the model to maintain normal classification performance on clean inputs while triggering malicious behavior under specific input conditions. The core training objective can be formalized as a joint optimization problem that balances normal classification accuracy and backdoor behavior injection.

Let D denote the original clean dataset. A randomly selected subset of clean samples is denoted as D_2 and used to generate poisoned samples, while the remaining clean samples are denoted as D_1 and kept unmodified. Let y represent the true label of a sample, and y_{tar} denote the specified target label. For all samples in subset D_2, the following formula is applied:

$$x' = G(x, \text{Trigger}) \tag{3}$$

A trigger pattern is added to the sample using a generative function, denoted as the trigger-insertion function. A clean sample $(x, y) \in D_2$ is transformed via Equation (3), and its label y is altered to y_{tar} resulting in a malicious sample (x', y_{tar}). Repeating this procedure for all samples in D_2 yields a malicious subset D_2'. The poisoned dataset D_{poi} is then formed by mixing D_1 and D_2'. The prepared poisoned dataset D_{poi} is fed into a model parameterized by θ, with Loss(.) representing the loss function, such as the log-loss function. The overall training objective can be formulated as:

$$L = \min_{\theta}(L_1 + L_2) \tag{4}$$

where L_1 denotes the classification loss of the model on the clean subset, and L_2 represents the backdoor injection loss on the malicious subset, ensuring that the poisoned samples trigger the target output. The two loss components are defined as follows:

$$\mathrm{L_1} = \frac{1}{|\mathrm{D_1}|} \sum_{(\mathrm{x,y})\in \mathrm{D_1}} \mathrm{Loss}(\mathrm{f_\theta(x)}, \mathrm{y}) \tag{5}$$

$$L_2 = \frac{1}{|D_2^{'}|} \sum_{(x^{'},y)\in D_2^{'}} \mathrm{Loss}(f_\theta(x^{'}), y_{\mathrm{tar}}) \tag{6}$$

However, existing studies have revealed that poisoned samples often exhibit notable clustering shifts in deep feature spaces, a phenomenon known as latent separability, making them susceptible to detection through feature-space analysis methods, such as activation clustering or spectral signatures. To circumvent such detection mechanisms, this paper introduces a feature similarity regularization term (FSR) during training. The intuition behind using the mean squared error lies in treating the feature representations of both poisoned and clean samples as a unified cluster. This term penalizes the mean squared distance of all sample points from the cluster center, thereby constraining the model to learn representations where poisoned samples lie closer to the center of clean feature distribution. Let z denote the mapping function from input to feature space and let $z(x)$ denote the feature representation extracted from the penultimate layer, and let $\bar{z}$ denote the feature centroid of clean samples in D_1.

$$\bar{z} = \frac{1}{|D_1|} \sum_{(x,y)\in D_1} z(x) \tag{7}$$

The regularization loss term can be formulated as:

$$L_3 = \frac{1}{|D_2'|} \sum_{(x',y_{tar})\in D_2'} ||z(x') - \bar{z}||_2^2 \tag{8}$$

The final total training loss function can be expressed as:

$$\min_{\theta} L = L_1 + L_2 + \lambda L_3 \tag{9}$$

where λ denotes the balancing weight of the feature similarity regularization term.

Through the above joint optimization design, the model not only learns the backdoor behavior but also actively guides the poisoned samples to embed into the distribution cluster of benign samples in the feature space. Thereby, it reduces their separability and enhances the ability to evade feature-space detection algorithms.

3.3. Attack framework design and overall procedure

The overall framework of the FBFS method is illustrated in Figure 4. Its core procedure can be divided into two stages: sample construction and model training.

(1) Poisoned Dataset Construction

First, we adopt an all-to-one attack strategy, designating the “balloon” class as the target. Before model training, a randomly selected subset of Red, Green, and Blue (RGB) clean images S (of size

3 × H × W is preprocessed (e.g., cropped and resized to a uniform size). A physical fog model is then applied to these images to simulate fog effects, while their labels are changed to "balloon". These modified samples are mixed with the remaining unaltered clean images to form the poisoned dataset.

(2) Joint Optimization Training

The constructed poisoned dataset is fed into the model, and the objective function is modified to guide the model in learning backdoor features. Throughout this process, the model employs multiple convolutional layers to extract feature vectors from the images. These features are further processed by fully connected layers to combine discriminative characteristics and learn representations of different classes.

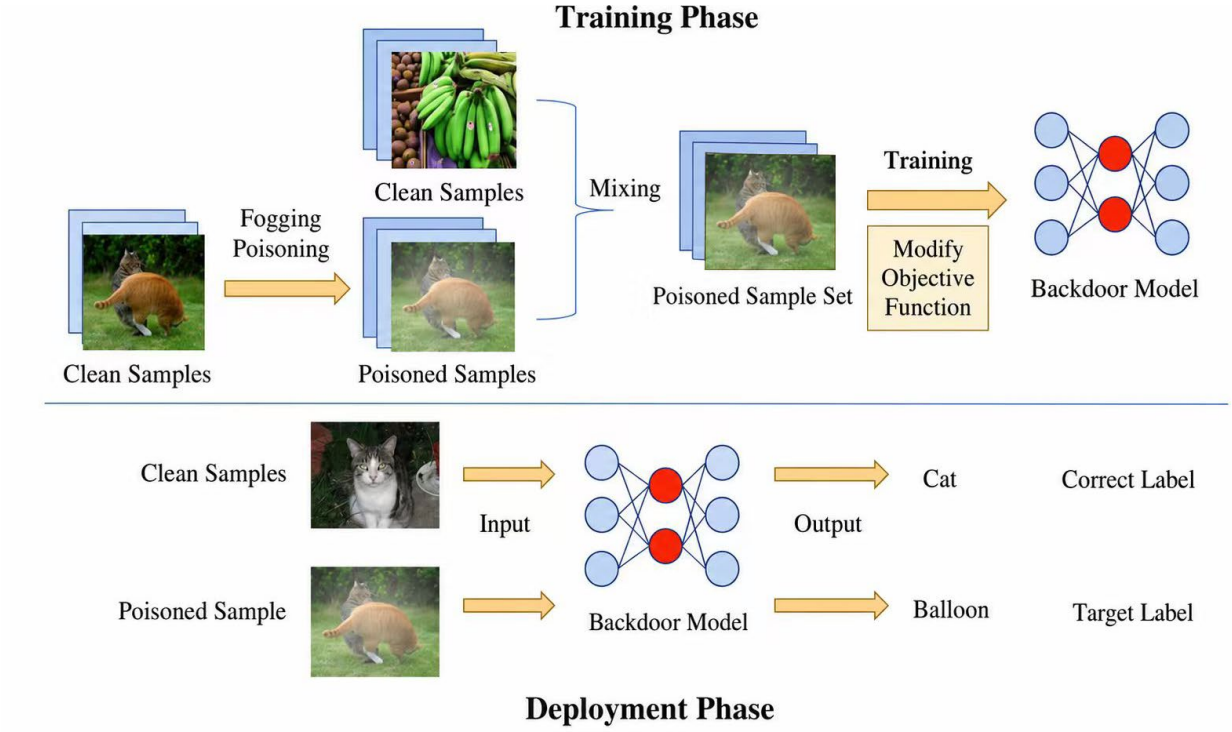

Figure 4. Flowchart of the principle of the fogging backdoor attack model.

During training, a joint loss function as defined in Equation (9) is optimized, which consists of the standard classification loss, the backdoor injection loss, and the feature similarity regularization term.

By incorporating the feature similarity constraint, the distribution gap between poisoned and clean samples in the deep feature space is actively reduced. This effectively mitigates the latent separability of backdoor samples and enhances the evasion capability against feature-space detection mechanisms.

The FBFS method does not rely on specific network architectures or training frameworks, offering both simplicity and transferability. It is applicable to backdoor attacks across various class scales and task scenarios. Without modifying the model structure or significantly increasing training complexity, the method achieves improved visual stealth and stronger resistance to detection.

3.4. Pseudo-code of the FBFS training procedure

The proposed FBFS algorithm is summarized in pseudo-code as Algorithm 1.

Algorithm 1 Fog-based Backdoor Attack with Feature Similarity Regularization

Input: Clean sample pairs(x, y)∈D1, poisoned sample pairs (x', yt)∈D2', poisoned dataset Dpoi model parameters θi per batch.

Output: Final model parameters θ.

1. Initialize model parameters θ;
2. Repeat:
3. for epoch i = 1to N do
4. Compute and update the loss on clean training subset using Equation (5);
5. Compute and update the loss on poisoned training subset using Equation (6);
6. Compute and update the feature-space regularization term using Equation (8);
7. Calculate the total loss using Equation (9):

$L_i = L_{i1} + L_{i2} + \lambda L_{i3}$;

8. Update model parameters via gradient descent:

$$\theta_{i+1} = \theta_i - \eta \nabla L_i.$$

9. epoch = epoch + 1

4. Experiments and analysis

To evaluate the performance of FBFS, experiments were conducted on three datasets and compared with two baseline attack methods. The effectiveness of FBFS was validated, and the robustness of the three attacks against specific defense mechanisms was compared.

4.1. Experimental settings and evaluation metrics

(1) Datasets and Model Configuration

Three widely used datasets in backdoor attack and defense research were selected for fog-based backdoor attack experiments: CIFAR-10 [28], GTSRB [29], ImageNet [30]. CIFAR-10 and GTSRB are commonly adopted in backdoor research. In particular, GTSRB clearly demonstrates the high risks of backdoor attacks in autonomous driving scenarios. Due to computational constraints, a subset of 10 classes from ImageNet was used, including: banana, baboon, kayak, balloon, orange, cat, table, mug, soccer ball, dumbbell. The training set consists of 900 images per class (9000 in total), and the test set contains 1000 images, resulting in a dataset of 10,000 images. Sample images from the three datasets are shown in Figure 5 and detailed dataset statistics are provided in Table 1.

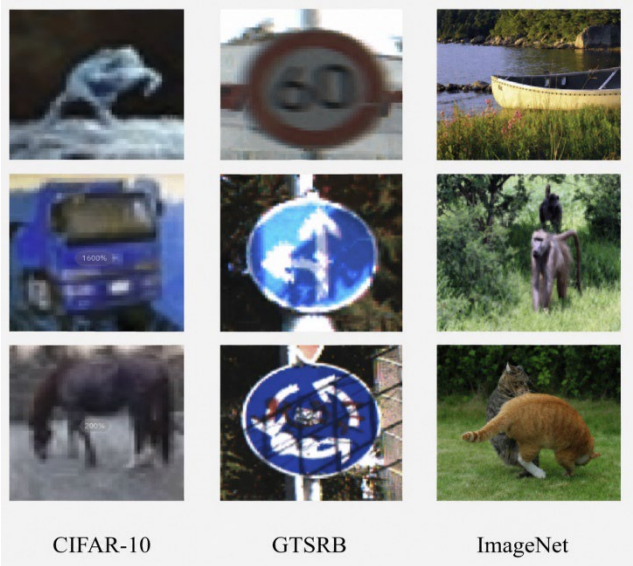

Figure 5. Some image examples from the dataset in this paper.

Table 1. Experimental datasets.

Data	Class	Input Size	Trainset Size
CIFAR-10	10	32 × 32 × 3	50,000
GTSRB	43	32 × 32 × 3	39,209
ImageNet Subset	10	224 × 224 × 3	9000

For model architecture, the ResNet [31] series was uniformly adopted due to its strong feature extraction capability. Specifically, ResNet-18 pre-trained models were used for CIFAR-10 and GTSRB. ResNet-34 pre-trained models were used for the ImageNet subset. All models were initialized using the Xavier initialization method to facilitate faster convergence and stable performance. The gradients during parameter propagation were optimized using the Stochastic Gradient Descent (SGD) algorithm.

All experiments were implemented using the PyTorch framework. The hardware configuration included an NVIDIA GeForce RTX 3090 GPU, an Intel(R) Xeon(R) Platinum 8352 CPU with a base frequency of 2.10 GHz, 240 GB of RAM, and Python as the programming language. The models were trained for 200 epochs. SGD [32] was used for gradient updates, with a StepLR learning rate scheduler applied at fixed intervals. The coefficient λ for the regularization term in the objective function was set to $\lambda = 0.2$.

(2) Baseline Methods

To demonstrate the performance advantages of FBFS, two representative backdoor attack methods were selected for comparison:

BadNets [1]: A classic fixed-pattern backdoor attack.

Adapted Patched [9]: A representative feature-space suppression-based backdoor attack.

The poisoning rate for all attacks was set to 10%.

(3) Defense and Detection Methods

To verify whether the proposed method effectively suppresses the heterogeneous distribution characteristics of backdoor samples in the feature space, three defense methods were employed: Spectral Signature [24], Activation Clustering [23], and FEAT-IN [33]. All three defenses are based on detecting heterogeneous feature distributions of backdoor samples.

(4) Evaluation Metrics

For a clean classification model, classification accuracy (ACC) is sufficient to evaluate performance.

For a backdoored model, it is necessary to evaluate both: the model's performance under normal operation (when the backdoor is not triggered), and the success rate of malicious behavior when the backdoor is activated.

Therefore, two metrics are used: Clean Accuracy (CA) and Attack Success Rate (ASR) are defined in Equations (10)–(11). CA serves as a crucial metric for evaluating the stealthiness of backdoor attacks. When compared with the Accuracy (ACC) of the corresponding clean model, it effectively assesses the impact of the implanted backdoor on the model's classification performance. The ASR is used to measure the effectiveness of the backdoor attack, specifically quantifying the model's ability to misclassify triggered samples into the target category.

$$\mathrm{CA} = \frac{1}{|D_1|} \sum_{(x,y)\in D_1} I(f_\theta(x) = y) \tag{10}$$

$$\mathrm{ASR} = \frac{1}{|D_2'|} \sum_{(x',y)\in y_{\mathrm{tar}}} I(f_\theta(x') = y_{\mathrm{tar}}) \tag{11}$$

To evaluate the capability of the backdoor attack algorithm in evading backdoor defense methods, this paper adopts the metrics proposed in [34]:

Elimination Rate (ER): The proportion of successfully detected poisoned samples.

Sacrifice Rate (SR): The proportion of clean samples incorrectly identified as poisoned.

Let Def denote the defense method and S represent the total number of samples predicted as poisoned. ER and SR can be expressed by Equations (12) and (13), respectively. These two metrics also serve as important indicators for assessing the performance of the backdoor attack algorithm.

$$\mathrm{ER} = \frac{1}{|\mathrm{D}_2'|} \sum_{(\mathrm{x}',\mathrm{y})\in \mathrm{D}_2'} \mathrm{I}(\mathrm{Def}(\mathrm{x}') = \mathrm{y}_{\mathrm{tar}}) \tag{12}$$

$$\mathrm{SR} = \frac{1}{|S|} \sum_{(x,y)\in D_1} I(\mathrm{Def}(x) = y_{\mathrm{tar}}) \tag{13}$$

4.2. Attack effectiveness analysis

We first evaluate the effectiveness of the proposed FBFS method under no defense mechanisms across different datasets and compare it with existing classical backdoor attack methods. Table 2 presents the experimental results of BadNets, Adapted Patched, and FBFS under the same poisoning rate (10%) on CIFAR-10, GTSRB, and the ImageNet subset.

Table 2. Attack performance comparison under 10% poisoning rate.

Datasets	Attack methods	Accuracy	ASR	CA
CIFAR-10	BadNets [1]	97.52	**98.67**	**97.70**
	Adapted Patched [9]	97.52	85.55	88.55
	FBFS (Ours)	97.52	93.83	95.41
GTSRB	BadNets [1]	95.21	**94.48**	**91.09**
	Adapted Patched [9]	95.21	84.78	87.78
	FBFS (Ours)	95.21	89.43	90.98
ImageNet	BadNets [1]	91.63	**95.34**	**90.18**
	Adapted Patched [9]	91.63	83.20	85.37
	FBFS (Ours)	91.63	86.38	88.72

The results demonstrate that the proposed FBFS method achieves comparable or even higher ASR than existing methods across all datasets. On both CIFAR-10 and GTSRB, FBFS attains an ASR exceeding 92%, while on the ImageNet subset, it maintains an ASR above 90%, indicating strong cross-dataset adaptability and attack stability. Furthermore, CA of FBFS remains above 85% in all cases,

confirming that the injection of the backdoor does not significantly degrade the model's performance on benign samples.

It is worth noting that while the Adapted Patched method also achieves high ASR in certain tasks, its training process is considerably more complex than that of FBFS. Although BadNets is easier to implement in some scenarios, its ASR exhibits noticeable instability on more complex datasets such as ImageNet. Overall, FBFS achieves high attack success rates while offering greater training efficiency and broader applicability.

4.3. Defense robustness evaluation

To verify the effectiveness of FBFS in evading feature-space detection mechanisms, we evaluate its performance against three backdoor detection methods: Activation Clustering (AC), Spectral Signature (SS), and the recently proposed FEAT-IN. These detection approaches are all based on the separability of poisoned samples in the feature space and are known for their strong detection capabilities.

Table 3 summarizes the detection results of the three backdoor attack methods under the three detection mechanisms across CIFAR-10, GTSRB, and the ImageNet subset. Key observations include:

(1) BadNets exhibits high detection rates under both AC and SS, especially on CIFAR-10, where the ER consistently exceeds 85%, with a low SR below 19%. This indicates that traditional backdoor attacks exhibit noticeable anomalies in feature distributions.

(2) The Adapted Patched method, which uses an asymmetric trigger design, partially suppresses feature-space separability and performs significantly better than BadNets under AC and SS. However, it still faces a high detection risk under FEAT-IN, with an ER exceeding 70%.

(3) In contrast, FBFS demonstrates considerably lower detection rates across all detection mechanisms. On CIFAR-10, GTSRB, and ImageNet, the ER of FBFS remains below 22% under both AC and SS, while the SR is also kept at a low level.

Table 3. Defense experiment results of the proposed model.

Datasets	Defend methods Attack method	Spectral Signature [24]		Activation Clustering [23]		FEAT-IN [33]	
		ER	SR	ER	SR	ER	SR
CIFAR-10	BadNets [1]	98.0	4.2	100.0	7.1	100.0	5.9
	Adapted Patched [9]	**15.3**	18.4	16.7	18.8	86.3	33.8
	FBFS (Ours)	19.6	**24.8**	**15.1**	**20.7**	**44.6**	**52.4**
GTSRB	BadNets [1]	95.2	8.3	91.1	18.7	97.4	8.2
	Adapted Patched [9]	22.2	25.7	23.7	**20.2**	74.2	29.3
	FBFS (Ours)	**17.2**	**26.8**	**18.9**	17.2	**51.6**	**41.3**
ImageNet	BadNets [1]	86.2	11.3	89.3	12.4	93.3	10.7
	Adapted Patched [9]	**19.8**	**27.3**	**17.4**	**24.9**	76.1	31.7
	FBFS (Ours)	20.6	16.3	21.7	18.8	**42.5**	**47.7**

Under the more advanced FEAT-IN detection method, BadNets is inevitably identified with an ER over 70%, and even with an SR of approximately 30%, it fails to evade detection effectively. Although FBFS shows detection performance similar to Adapted Patched under FEAT-IN, it exhibits greater overall stability. While FEAT-IN can detect approximately half of the backdoor samples injected by FBFS, it comes at the cost of an SR close to 50%, implying that a large number of clean samples are

misclassified as malicious. This limited detection efficacy further validates the effectiveness and robustness of the feature-space heterogeneity suppression strategy adopted in FBFS.

The above results demonstrate that FBFS effectively reduces the distribution gap between poisoned and clean samples in the feature space through feature similarity regularization, thereby weakening the separability assumption relied upon by existing detection methods. As a result, FBFS achieves high attack success rates while significantly enhancing the model's evasion capability against detection.

4.4. Feature-space visualization analysis

To further validate the superiority of FBFS in feature-space blending, we employ t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the deep feature representations under different backdoor attack methods. Feature vectors extracted from the penultimate layer of the model for both clean and poisoned samples are projected into a 2D space to observe their distribution patterns.

Figure 6 illustrates the feature-space visualization results for BadNets, Adapted Patched, and FBFS on the CIFAR-10 dataset. Key observations include:

(1) Under BadNets, poisoned and clean samples form clearly separable clusters in the feature space, with poisoned samples aggregating into distinct isolated clusters that are easily detectable by clustering-based detection algorithms.

(2) Adapted Patched partially mitigates the clustering shift of poisoned samples, but certain regions still exhibit discernible boundaries.

(3) By incorporating feature similarity regularization, FBFS effectively compresses the feature deviation of poisoned samples. Their distribution is largely blended within the clusters of normal samples, significantly reducing overall separability.

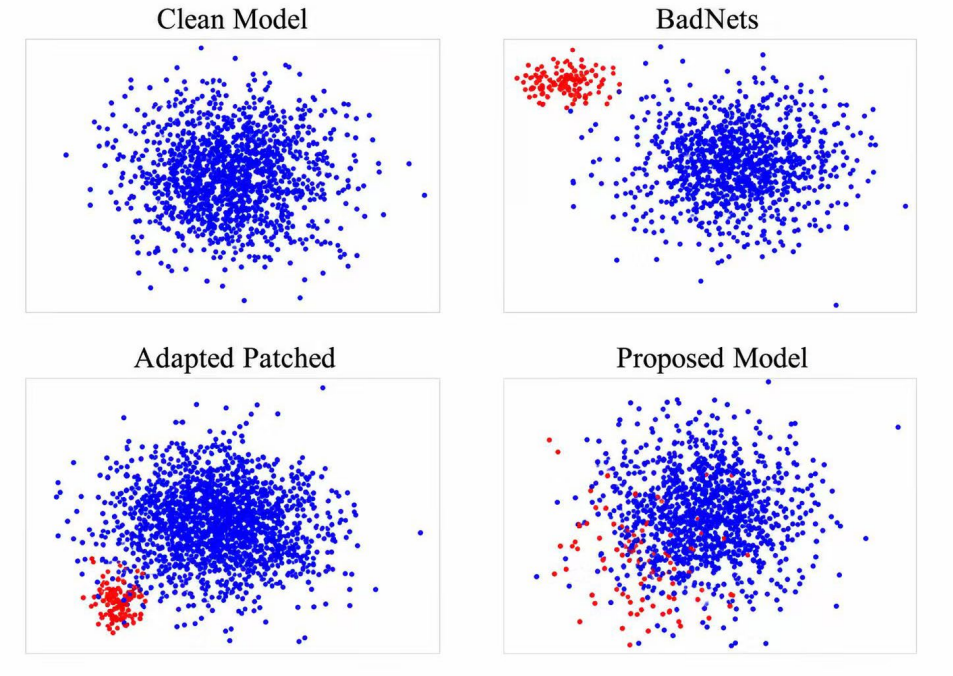

Figure 6. t-SNE visualization across different models.

These visualization results strongly indicate that FBFS successfully suppresses the anomalous distribution characteristics of poisoned samples in the deep feature space, providing robust support for evading detection methods based on latent separability.

5. Ablation study and parameter analysis

5.1. Analysis of fog generation parameters

In the construction of the fog-based trigger, the values of the atmospheric light parameter α and the scattering coefficient β in the standard optical model directly affect the brightness and density of the fog. To ensure the natural appearance and stealthiness of the backdoor trigger, we first conduct systematic experiments tuning these two parameters. Here, α primarily controls overall brightness, while β regulates fog concentration.

Empirical observations show that when $\beta > 0.2$, fog tends to form noticeable patches, and high fog density often causes image distortion. Therefore, we set $\beta \in \{0.05, 0.10, 0.15, 0.20\}$ and correspondingly choose $\alpha \in \{0.50, 0.55, 0.60, 0.65\}$, maintaining a linear relationship between them to simulate natural fog effects. Values of $\alpha < 0.5$ were not considered, as they produce unnatural grayish-black fog effects.

Figure 7 displays fog effects under different parameter combinations. It can be observed that as both parameters increase, the fog effect becomes more pronounced.

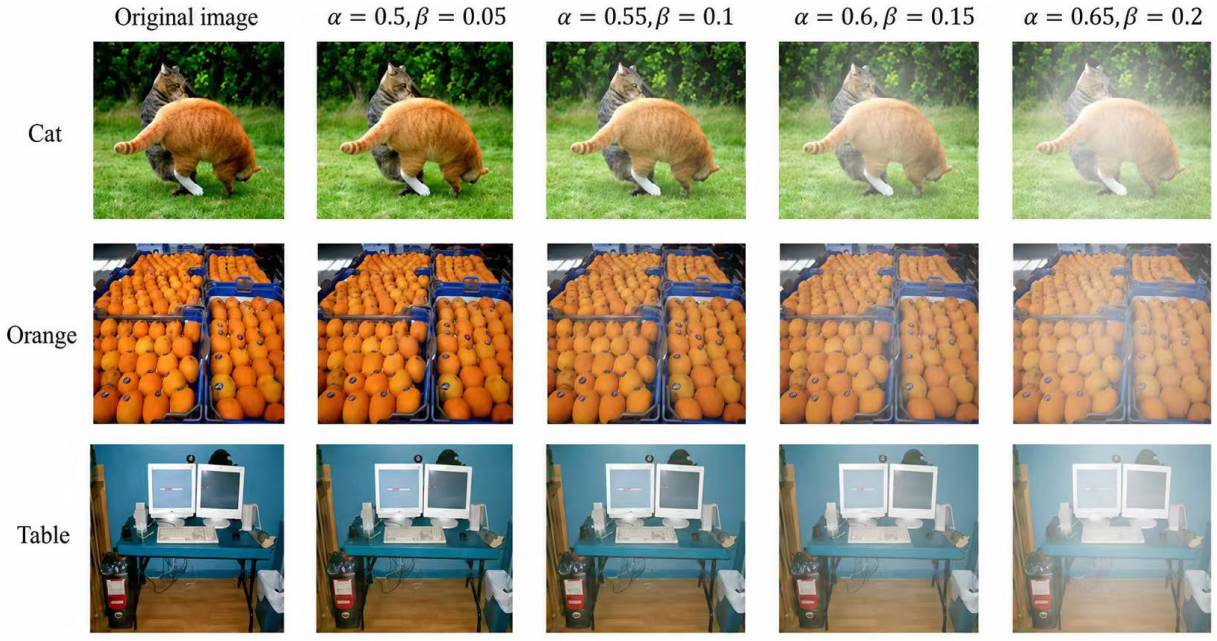

Figure 7. Foggy image samples under various parameters.

To further quantify the impact of different fog parameters on backdoor attack performance and image quality, this study evaluates the ASR, CA, and Structural Similarity Index (SSIM) under various parameter combinations using the ResNet-34 model on an ImageNet subset. The SGD optimizer was employed over 200 training epochs with a StepLR learning rate scheduler. The regularization coefficient was set to 0.2, and model parameters were initialized using the Xavier method to accelerate convergence. The SSIM metric, which incorporates luminance, contrast, and structural information, provides a perceptually relevant measure of image quality.

As shown in the left panel of Figure 8, the parameter set $\alpha = 0.55$, $\beta = 0.10$ achieves the best balance between ASR and CA, maintaining a high ASR while keeping CA close to that of the clean model. The right panel indicates that this configuration also yields an SSIM value close to 0.9, demonstrating that an effective trigger is embedded without significant degradation of image quality. Based on these results, $\alpha = 0.55$ and $\beta = 0.10$ are selected as the default fog parameters for all subsequent experiments.

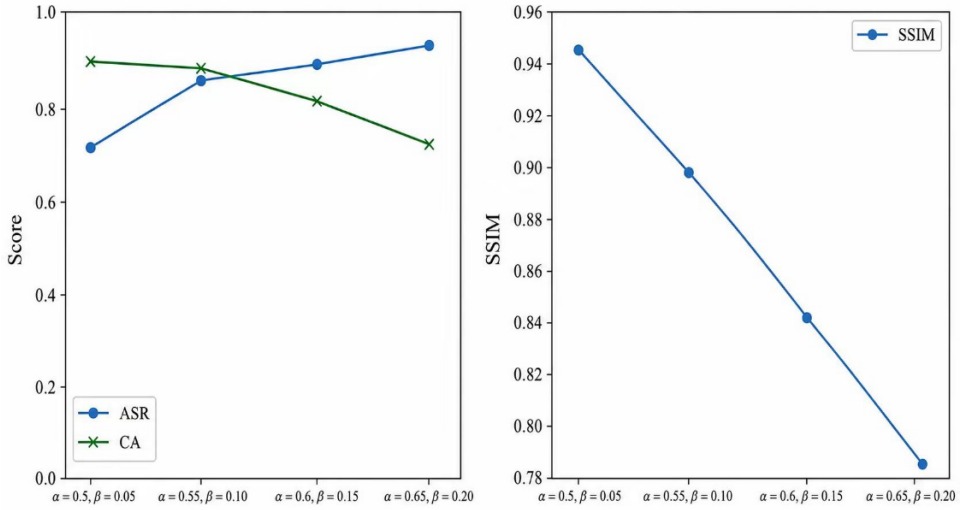

Figure 8. Impact of parameter pairs on attack performance and SSIM.

5.2. Poisoning rate sensitivity analysis

To evaluate the impact of the poisoning ratio on the performance of the FBFS attack, experiments were conducted on an ImageNet subset with poisoning rates ρ set at six levels: [1%, 3%, 5%, 7%, 10%, 15%]. The corresponding changes in ASR and CA were measured, and the results are shown in Figure 9.

As observed in the left panel of Figure 9, the ASR of the model increases consistently with higher poisoning rates. When the poisoning rate reaches 7%, the ASR exceeds 80%, demonstrating high attack effectiveness. In the right panel, the CA shows a slight decline as the poisoning rate increases, with a more pronounced degradation in classification performance observed when the poisoning rate exceeds 10%. Considering the trade-off between ASR and CA, a default poisoning rate of 10% is selected to ensure stable attack performance while preserving model utility.

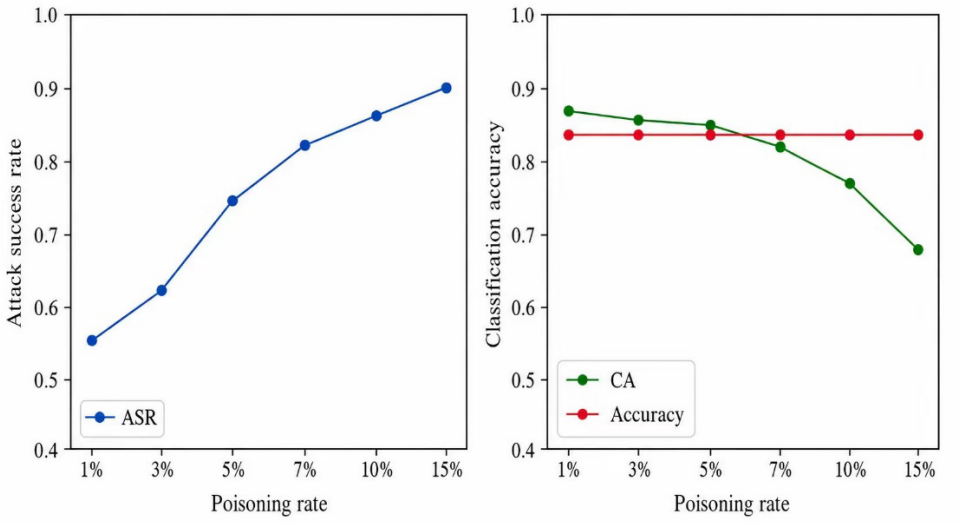

Figure 9. Effect of poisoning rates on attack performance.

6. Conclusions

This paper addresses the trade-off between visual stealth and resistance to feature-space detection in existing backdoor attacks by proposing FBFS. The method employs a standard optical model to generate natural fog as a visual trigger, enhancing both semantic plausibility and visual camouflage in poisoned images. Furthermore, a feature similarity regularization term is introduced during training to effectively reduce the distribution discrepancy between poisoned and clean samples in the deep feature space, thereby suppressing their latent separability and evading feature-space detection mechanisms.

Extensive experiments on CIFAR-10, GTSRB, and a subset of ImageNet demonstrate that FBFS maintains the model's clean classification accuracy while achieving higher attack success rates and lower detection rates. Compared to existing representative backdoor methods, FBFS exhibits superior performance in both attack effectiveness and evasion capability. Feature-space visualizations further confirm its effectiveness in blending feature distributions.

Future work will focus on extending and optimizing FBFS for object detection, cross-modal tasks, and dynamic triggering strategies, with the goal of further improving its stealth and adaptive capabilities in complex scenarios.

Data availability declaration

The datasets analyzed in this study are publicly available. The CIFAR-10 dataset is available at https://www.cs.toronto.edu/~kriz/cifar.html; the German Traffic Sign Recognition Benchmark (GTSRB) dataset is available at https://benchmark.ini.rub.de/gtsrb_dataset.html; and the ImageNet dataset is available at https://www.image-net.org/download.php, subject to ImageNet's registration and access conditions. No new primary datasets were generated in this study.

Declaration of generative AI and AI-assisted technologies

The authors did not use generative AI or AI-assisted technologies in the writing of this manuscript.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China Nos. 61902164 and 62462053, in part by the Key Research and Development Program of Gansu Province No. 24YFGA004.

Authors' contribution

Conceptualization, B.Z. and X.J.; methodology, B.Z. and W.Z.; software, B.Z. and W.Z.; validation, W.Z., H.Y. and G.Y.; formal analysis, B.Z.; investigation, B.Z. and H.Y.; resources, G.Y. and X.J.; data curation, W.Z.; writing—original draft preparation, B.Z.; writing—review and editing, W.Z., H.Y., G.Y. and X.J.; visualization, H.Y.; supervision, X.J.; project administration, X.J.; funding acquisition, G.Y. and X.J. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Gu T, Dolan-Gavitt B, Garg S. BadNets: identifying vulnerabilities in the machine learning model supply chain. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, Dallas, USA, November 3, 2017, pp. 1–11.

[2] Zhang J, Zhu C, Cheng X, Sun X, Chen B. Backdoor defense method in federated learning based on contrastive training. *J. Commun*. 2024, 45(3):182–196.

[3] Zhang Q, Yu M, Wang R, Li Y, Yuan J, *et al.* A sparse and invisible targeted backdoor attack in federated learning. *J. King Saud Univ. Comput. Inf. Sci.* 2025, 37:134.

[4] Xu W, Xu Y, Zhang S. Sample-independent federated learning backdoor attack in speaker recognition. *Cluster Comput*. 2025, 28(3):158.

[5] Chen J, Xiong H, Ma H, Zheng Y. CLB-Defense: based on contrastive learning defense for graph neural network against backdoor attack. *J. Commun*. 2023, 44(4):154–166.

[6] Chen Y, Bin Z, Zhao H. Stealthy graph backdoor attack based on feature trigger. *Complex Intell. Syst.* 2025, 11:325.

[7] Ma B, Guo Y, Ma J, Zhang Q, Fang C. Escape method of malicious traffic based on backdoor attack. *J. Commun*. 2024, 45(4):73–83.

[8] Han Y, Zhao B, Chu R, Luo F, Sikdar B, *et al*. UIBDiffusion: universal imperceptible backdoor attack for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, June 11–15, 2025, pp. 19186–19196.

[9] Qi X, Xie T, Li Y, Mahloujifar S, Mitta P. Revisiting the assumption of latent separability for backdoor defenses. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, May 1–5, 2023.

[10] Zhong H, Liao C, Squicciarini AC, Zhu S, Miller D. Backdoor embedding in convolutional neural network models via invisible perturbation. In *Proceedings of the ACM Conference on Data and Application Security and Privacy*, New Orleans, USA, March 16–18, 2020, pp. 97–108.

[11] Liu Y, Ma X, Bailey J, Lu F. Reflection backdoor: a natural backdoor attack on deep neural networks. In *Proceedings of the European Conference on Computer Vision*, Glasgow, UK, August 23–28, 2020, pp. 182–199.

[12] Xue M, Ni S, Wu Y, Zhang Y, Wang J, *et al*. Imperceptible and multi-channel backdoor attack. *Appl. Intell.* 2024, 54(3):1099–1116.

[13] Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks. *arXiv* 2019, arXiv:1912.02771.

[14] Mo J, Xu M, Xing X. Clean-label backdoor attack on link prediction task. *Cybersecurity* 2025, 8(1):51.

[15] Li Y, Li Y, Wu B, Li L, He R, *et al.* Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, October 11–17, 2021, pp. 16463–16472.

[16] Hou Y, Ying Y. Invisible backdoor attack with image contours triggers. In *Proceedings of Algorithms and Architectures for Parallel Processing*, Macau, China, October 29–31, 2024, pp. 134–153.

[17] Gong X, Tian B, Xue M, Li S, Chen Y, *et al*. Megatron: evasive clean-label backdoor attacks against vision transformer. *arXiv* 2024, arXiv:2412.04776.

[18] Dai E, Lin M, Zhang X, Wang S. Unnoticeable backdoor attacks on graph neural networks. In *Proceedings of the ACM Web Conference*, Austin, USA, April 30–May 4, 2023, pp. 2263–2273.

[19] Feng L, Qian Z, Li S, Zhang X. Trojaning semi-supervised learning model via poisoning wild images on the web. *arXiv* 2023, arXiv:2301.00435.

[20] Guo Y, Zhong N, Qian Z, Zhang X. Physical invisible backdoor based on camera imaging. *arXiv* 2023, arXiv:2309.07428.

[21] Li M, Zhong N, Zhang X, Qian Z, Li S. Object-oriented backdoor attack against image captioning. *arXiv* 2024, arXiv:2401.02600.

[22] Wang B, Yao Y, Shan S, Li H. Neural cleanse: identifying and mitigating backdoor attacks in neural networks. In *Proceedings of IEEE Symposium on Security and Privacy*, San Francisco, USA, May 19–23, 2019, pp. 707–723.

[23] Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, *et al.* Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv* 2018, arXiv:1811.03728.

[24] Tran B, Li J, Madry A. Spectral signatures in backdoor attacks. Advances in neural information processing systems. *arXiv* 2018, arXiv:1811.00636.

[25] Guo W, Demontis A, Pintor M, Chan PPK, Biggio B. LFPD: local-feature-powered defense against adaptive backdoor attacks. In *Proceedings of the 2024 International Conference on Machine Learning and Cybernetics*, Miyazaki, Japan, September 20–23, 2024, pp. 607–612.

[26] Zhou R, Meng S, Li M, Qiu S. Quantitative simulation method of aerodrome fog images based on atmospheric scattering model. *Ship Electron. Eng.* 2024, 44(1):99–104.

[27] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, *et al.* Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems*, Montreal, Canada, December 8–13, 2014, pp. 2672–2680.

[28] Krizhevsky A. Hinton G. Learning multiple layers of features from tiny images. Handbook of Systemic Autoimmune Diseases, 2009, pp. 1–60. Available: https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf (accessed on 18 May 2025).

[29] Stallkamp J, Schlipsing M, Salmen J, Igel C. Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks* 2012, 32(1):323–332.

[30] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, *et al.* Imagenet large scale visual recognition challenge. *Int. J. Comput Vision* 2015, 115(3):211–252.

[31] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, June 27–30, 2016, pp. 770–778.

[32] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res*. 2010, 9:249–256.

[33] Vršnak D, Subašić M, Lončarić S. Efficient method for robust backdoor detection and removal in feature space using clean data. *IEEE Access* 2025, 13:18215–18227.

[34] Liu Y, Ma S, Aafer Y, Lee WC, Zhai J, *et al.* Trojaning attack on neural networks. In *Proceedings of the Network and Distributed System Security Symposium*, California, USA, February 18–21, 2018, pp. 1–15.