Lite-BD: A Lightweight Black-box Backdoor Defense via Reviving Multi-Stage Image Transformations

Abdullah Arafat Miah*, Yu Bi*
* Department of Electrical, Computer, and Biomedical Engineering
University of Rhode Island, Kingston, RI, USA
{abdullaharafat.miah, yu_bi}@uri.edu

***Abstract*—Deep Neural Networks (DNNs) are vulnerable to backdoor attacks. Due to the nature of Machine Learning as a Service (MLaaS) applications, black-box defenses are more practical than white-box methods, yet existing purification techniques suffer from key limitations: a lack of justification for specific transformations, dataset dependency, high computational overhead, and a neglect of frequency-domain transformations. This paper conducts a preliminary study on various image transformations, identifying down-upscaling as the most effective backdoor trigger disruption technique. We subsequently propose `Lite-BD`, a lightweight two-stage blackbox backdoor defense. `Lite-BD` first employs a super-resolution-based down-upscaling stage to neutralize spatial triggers. A secondary stage utilizes query-based band-by-band frequency filtering to remove triggers hidden in specific bands. Extensive experiments against state-of-the-art attacks demonstrate that `Lite-BD` provides robust and efficient protection. Codes can be found at https://github.com/SiSL-URI/Lite-BD.**

***Index Terms*—Backdoor attacks, Trustworthy AI, Deep Neural Networks.**

I. Introduction

With the rise of Machine Learning as a Service (MLaaS) [8], Artificial Intelligence (AI) usage has become increasingly dependent on third parties, creating serious security vulnerabilities for end users, such as backdoor attacks [15]. In a conventional backdoor attack scenario, an adversary inserts a backdoor into a deep model by poisoning it with a malicious dataset, either during training from scratch or during fine-tuning at any stage of the supply chain. Consequently, the model performs normally when clean samples are fed into it but will predict an adversarial target when trigger patterns appear in the inputs. To defend against backdoor attacks, several defense methods have been proposed; however, most require white-box access to the model [13] [30]. While these methods can produce effective mitigation, their utility is limited for MLaaS systems where the end user lacks access to model parameters. To address this gap, researchers have proposed black-box backdoor defenses—such as backdoor sample detection [19], model detection [33], and sample purification [23], [34], where the defender has no access to the model internals. Despite the promises of sample purification based methods, they face several limitations: *i) there is often no proper justification for the choice of specific spatial transformations, ii) heavily adopted diffusion models may require costly task-specific training introducing significant computational overhead, and iii) if spatial transformations fail to disrupt a specific trigger, the potential for frequency filtering as an alternative remains largely unexplored.*

To this end, we first evaluate various spatial transformations through a preliminary study, identifying downscaling-then-upscaling as the most effective trigger disruption method. We then propose **`Lite-BD`**, a novel and lightweight black-box trigger purification method. In the upscaling phase, we employ neural super-resolution [28], [17] to provide effective image restoration without the high computational cost of diffusion models. Our zero-shot approach eliminates dataset dependency and training overhead. For triggers that bypass spatial

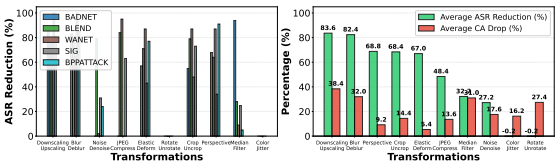

Fig. 1. Spatial transformation effects on backdoor attacks. The left panel shows ASR reduction across all five attacks and transformations. The right panel presents average performance metrics sorted by ASR reduction.

transformations, we further propose a band-by-band frequency filtering technique to neutralize signatures in the frequency domain. In summary, this paper makes the following three main contributions:

- We evaluate various spatial and inverse transformations to identify optimal configurations for disrupting backdoor triggers.
- We propose **`Lite-BD`**, a novel two-stage black-box backdoor defense framework. The first stage purifies poisoned samples through a pipeline of downscaling and neural super-resolution-based upscaling; if the trigger persists, a second stage employs query-based band-by-band frequency filtering to neutralize triggers in the frequency domain.
- Through extensive experiments, we demonstrate that **`Lite-BD`** effectively reduces Attack Success Rate (ASR) with minimal impact on benign performance, outperforming existing black-box methods in computational efficiency.

II. Related Works

Backdoor Attacks. Existing backdoor attacks can be broadly categorized into those using visible and invisible triggers. For example, BadNets [8] employs patch-based visible trigger patterns, whereas Blend [2] and WaNet [22] utilize invisible triggers. Triggers can also be adaptive, such as LIRA [3], or image-quantization-based, as in BppAttack [29]. Backdoor attacks may operate in the frequency domain, such as FIBA [5] and LF [16], or jointly in both the spatial and frequency domains, as demonstrated by DUBA [6]. More recently, attacks have been extended to simultaneously exploit the spatial, frequency, and feature domains, as in [7]. Additionally, some attacks can be mounted without label poisoning, including SIG [1] and LC [27].

Backdoor Defense. In white-box backdoor defense, the defender has access to the model and can control the training process [13], perform fine-tuning [37], [38], or apply pruning-based strategies [18], [30], [14]. In the black-box setting, backdoors can be mitigated through backdoored model detection [31], poisoned sample detection [19], or sample purification methods that aim to destroy triggers at test time, such as Sancdifi [21], BDMAE [26], ZIP [23], and SampDetox [34]. However, existing purification-based black-box defenses suffer from several key limitations, including effectiveness restricted to local triggers, strong dataset dependency, and additional computational overhead.

III. Preliminary Study

To evaluate the efficacy of spatial transformations in disrupting backdoor triggers, we conduct an initial study using five representative attacks: BadNets [8], Blend [2], WaNet [22], SIG [1], and BppAttack [29]. Using ResNet-18 on CIFAR-10, we measure Attack Success Rate (ASR) and Clean Accuracy (CA) across 100 random samples and their poisoned counterparts before and after ten distinct transformation-inverse pairs. Evaluated operations include: *Down-Upscaling* (50% bilinear), *Blur-Deblur* (5×5 Gaussian kernel/unsharp mask), *Noise-Denoise* (Gaussian noise/Non-Local Means), *JPEG Compression* (quality 50), *Elastic Deformation*, *Rotate-Unrotate* ($15°$), *Crop-Uncrop*, *Perspective Transformation*, *Median Filtering*, and *Color Jittering*. Results in Figure I indicate that *Downscaling-Upscaling* achieves the highest average ASR reduction, followed by *Blur-Deblur*. While ShrinkPad [15] utilizes image shrinking, its reliance on padding significantly degrades CA. Recent defenses like ZIP [23] and SampDetox [34] employ *Blur-Deblur* and *Noise-Denoise* respectively, but rely on computationally expensive diffusion models for restoration.

IV. Methodology

A. Threat Model

We assume a black-box defense scenario where the defender lacks access to model parameters

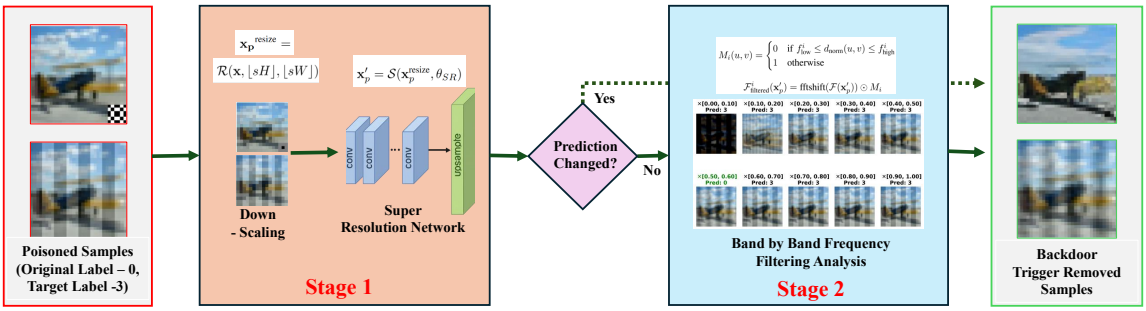

Fig. 2. Overview of the Proposed Black-box Backdoor Defense `Lite-BD`.

and/or fine-tuning capabilities, and is restricted to observing output predictions. The defender's objective is to filter poisoned inputs and purify samples to disrupt triggers while maintaining benign semantics. Conversely, we assume a strong adversary with full control over the training process, capable of injecting triggers that cause malicious misclassification while preserving normal performance on clean inputs.

B. Overview

Based on the observations from the preliminary study, we propose a two-stage black-box defense framework `Lite-BD` designed to disrupt backdoor triggers in poisoned samples and purify them for correct classification into their benign labels. In the first stage, we implement a "Downscaling-Upscaling" operation by employing stochastic resizing followed by neural super-resolution. If this operation fails to disrupt the trigger, a condition determined by auxiliary queries that fails to show a label flip, the output is passed to the second stage. The subsequent stage utilizes a band-by-band frequency filtering technique specifically designed to destroy triggers that remain robust against the initial resizing operations. If neither stage results in a label flip, the input image is deemed backdoor-free. An overview of our proposed `Lite-BD` is given in Figure 2.

C. Stage 1: Stochastic Resize and Neural Super-Resolution

For a given poisoned image $\mathbf{x}_p \in \mathbf{R}^{H\times W\times C}$, where H, W, and C represent the image dimensions, we apply a stochastic downscaling operation with a scaling factor $s \in [s_{\min}, s_{\max}]$, where $s_{\min}, s_{\max} \in (0,1)$. The dimensions of the downscaled image are computed via Equation 1, where $\mathcal{B}(\cdot)$ denotes bilinear interpolation. Such geometric transformation effectively disrupts most backdoor triggers, as trigger patterns often rely on precise pixel-level placement.

$$\mathbf{x_p}^{\text{resize}} = \mathcal{B}(\mathbf{x}, \lfloor sH \rfloor, \lfloor sW \rfloor) \tag{1}$$

For a benign sample $\mathbf{x}_c$ and its poisoned counterpart $\mathbf{x}_p$, the downscaling operation yields $\mathbf{x}_p^{\text{resize}}$. Our objective is to apply an inverse transformation to restore the image to its original dimensions $H \times W$. Let the reconstructed image be denoted as $\mathbf{x}'_p$. For ideal purification, the condition $\mathbf{x}'_p = \mathbf{x}_c$ should be satisfied. To achieve this, we employ a pretrained super-resolution (SR) network, denoted as $\mathcal{S}(\cdot)$, to perform the upscaling:

$$\mathbf{x}'_p = \mathcal{S}(\mathbf{x}_p^{\text{resize}}, \theta_{SR}) \tag{2}$$

where θ_{SR} represents the parameters of the pretrained backbone. Specifically, we utilize the Enhanced Super-Resolution Generative Adversarial Network (Real-ESRGAN [28]) with $4\times$ upscaling and the Swin Transformer-based architecture (SwinIR [17]) with $2\times$ upscaling which are referred to as **`Lite-BD (RE)`** and **`Lite-BD (SW)`**, respectively. During downscaling, aggregating source pixels disrupts the spatial coherence of pixel-level triggers. Pretrained super-resolution networks then reconstruct semantic information using learned natural image priors rather than local artifacts. Because trigger patterns are typically out-of-distribution

(OOD) for SR networks, they are not recovered during upscaling. Consequently, the purified image $\mathbf{x}'_p$ retains semantic integrity while the malicious trigger is neutralized.

D. Stage 2: Band-by-Band Frequency Filtering

Following Stage 1, the purification of a potentially poisoned sample is verified by querying the target model. If the backdoor trigger has been successfully disrupted, the predicted label should no longer be flipped to the target class. However, if the trigger remains intact, the sample will proceed to Stage 2, where we conduct a band-by-band frequency filtering analysis. For a given input image from Stage 1, $\mathbf{x}'_p \in \mathbf{R}^{C\times H\times W}$, we apply a 2D Discrete Fourier Transform (DFT) to each color channel via Equation 3, where (u, v) represents the frequency domain coordinates. The resulting coefficients are shifted to center the zero-frequency component (DC component). We then perform frequency normalization using Equation 4 to map distances from the center:

$$\mathcal{F}(\mathbf{x}'_p)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathbf{x}'_p(h,w)\cdot e^{-j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)} \tag{3}$$

$$d_{\text{norm}}(u,v) = \frac{\sqrt{(u-u_c)^2+(v-v_c)^2}}{\sqrt{u_c^2+v_c^2}} \tag{4}$$

The normalized frequency spectrum is partitioned into n uniformly spaced concentric bands, where each band B_i is defined by the range $[f^i_{\text{low}}, f^i_{\text{high}}]$. For each band, we apply a band-stop filter mask M_i as defined in Equation 5. The filtered frequency representation is obtained through element-wise multiplication (Hadamard product) in Equation 6, and the filtered image is reconstructed using the Inverse Fourier Transform in Equation 7, where $\mathcal{R}\{\cdot\}$ extracts the real component:

$$M_i(u,v) = \begin{cases} 0 & \text{if } f^i_{\text{low}} \le d_{\text{norm}}(u,v) \le f^i_{\text{high}} \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

$$\mathcal{F}^i_{\text{filtered}}(\mathbf{x}'_p) = \text{fftshift}(\mathcal{F}(\mathbf{x}'_p)) \odot M_i \tag{6}$$

$$\tilde{\mathbf{x}}'^i_p = \mathcal{R}\{\mathcal{F}^{-1}(\text{ifftshift}(\mathcal{F}^i_{\text{filtered}}(\mathbf{x}'_p)))\} \tag{7}$$

This process generates n filtered candidates $\{\tilde{\mathbf{x}}'^1_p, \ldots, \tilde{\mathbf{x}}'^n_p\}$ from a single image $\mathbf{x}'_p$. We pass each candidate through the model to detect restored labels, indicating the trigger's frequency band. If multiple bands disrupt the trigger, we prioritize higher frequencies to preserve semantic integrity. The optimal candidate then undergoes bilateral filtering or unsharp masking to restore edge sharpness. If both stages fail to correct the label flip, the sample is considered clean, and the original image is returned.

V. Experiments

A. Experimental Setup

Dataset and Model Settings To demonstrate the effectiveness of `Lite-BD`, we evaluate our approach on three widely used datasets: CIFAR-10 [12], Fashion-MNIST [32], GTSRB [25]. We employ two commonly used architectures, ResNet-18 [9] and VGG-11 [24]. Specifically, CIFAR-10 and Fashion-MNIST are classified using ResNet-18, while GTSRB is evaluated using VGG-11. We also extend our evaluation on CIFAR-10 using MobileNetV1 [10], DenseNet-121 [11], and Vision Transformer (ViT-Base) [4] models to demonstrate the generalizability of our method across diverse architectures.

Evaluation Metrics To assess the effectiveness of our method and enable fair baseline comparisons, we adopt three evaluation metrics: (1) *Clean Accuracy (CA)*, which measures the classification accuracy on clean samples after passing through the defense framework; (2) *Poisoned Accuracy (PA)*, which denotes the classification accuracy of poisoned samples after defense; and (3) *Attack Success Rate (ASR)*, which represents the proportion of poisoned samples that successfully trigger the backdoor attack. An effective defense is characterized by higher CA, higher PA, and lower ASR.

Attack Methods We evaluate our defense framework against ten state-of-the-art (SOTA) backdoor attacks, including BadNets [8], Blend [2], WaNet [22], SIG [1], Clean-Label (CL) [27], BPPAttack [29], Trojan [20], LF [35], Poison-Ink [36], and LIRA [3]. In Figure 3 poisoned samples from all the attacks with their corresponding triggers are shown.

Baseline Defense Methods We compare our proposed defense framework with three SOTA black-box defense methods: (1) *ShrinkPad* [15], which

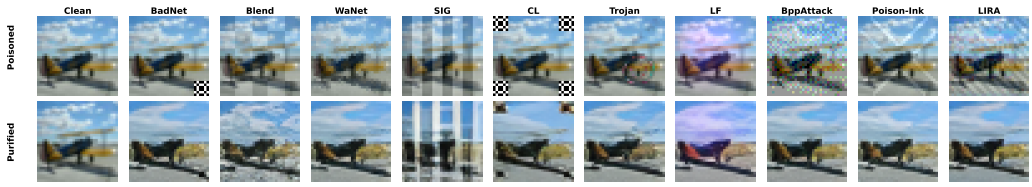

Fig. 3. Illustration of poisoned samples and their corresponding purified samples across ten backdoor attacks using our proposed method with `Lite-BD (RE)`.

TABLE I
DEFENSE PERFORMANCE COMPARISON. **BOLD** DENOTES THE BEST RESULT; UNDERLINE DENOTES THE SECOND-BEST.

Dataset	Attack	Base CA	Base ASR	ShrinkPad			ZIP			Sampdetox			Lite-BD (RE)			Lite-BD (SW)		
				CA	PA	ASR	CA	PA	ASR	CA	PA	ASR	CA	PA	ASR	CA	PA	ASR
CIFAR-10	BadNet	0.950	1.000	0.699	0.386	0.495	**0.886**	0.022	1.000	0.763	0.034	0.988	0.865	**0.911**	0.024	0.847	0.897	**0.024**
	Blend	0.946	0.999	0.731	0.350	0.299	0.881	0.724	0.064	0.772	0.648	0.063	**0.885**	**0.838**	**0.023**	0.861	0.659	0.164
	WaNet	0.945	1.000	0.556	0.395	0.592	0.879	0.827	0.054	0.788	0.737	0.054	**0.886**	**0.877**	0.052	0.870	0.857	**0.041**
	SIG	0.866	0.999	0.742	0.361	**0.001**	0.878	0.079	0.920	0.779	0.309	0.307	**0.888**	0.428	0.356	0.867	**0.516**	0.011
	CL	0.865	1.000	0.732	0.364	0.025	0.871	0.022	1.000	0.794	0.084	0.938	**0.889**	**0.879**	**0.000**	0.875	0.866	**0.000**
	BPPAttack	0.945	1.000	0.727	0.345	0.588	0.878	0.832	0.047	0.774	0.769	0.044	**0.881**	**0.903**	**0.013**	0.863	0.767	0.058
	Trojan	0.947	1.000	0.684	0.470	0.230	0.877	0.786	**0.047**	0.790	0.702	0.054	**0.881**	**0.806**	0.072	0.858	0.710	0.084
	LF	0.944	0.997	0.675	0.122	0.968	0.880	0.254	0.748	0.781	0.511	0.400	**0.884**	**0.893**	**0.016**	0.870	**0.893**	**0.016**
	Poison-Ink	0.946	0.998	0.677	0.294	0.535	**0.880**	**0.828**	0.057	0.769	0.753	**0.046**	0.873	0.779	0.148	0.864	0.631	0.209
	LIRA	0.941	1.000	0.738	0.517	0.170	0.875	0.746	0.102	0.783	0.736	0.042	**0.887**	**0.890**	**0.014**	0.852	0.567	0.224
	Average	0.933	0.999	0.696	0.360	0.390	0.879	0.512	0.404	0.779	0.498	0.294	**0.882**	**0.821**	**0.072**	0.863	0.736	0.083
Fashion-MNIST	BadNet	0.952	1.000	0.279	0.351	0.497	**0.897**	0.000	1.000	0.256	0.256	0.707	0.776	**0.807**	**0.038**	**0.845**	**0.807**	0.113
	Blend	0.950	0.999	0.268	0.266	0.164	**0.857**	0.684	0.191	0.837	**0.773**	**0.055**	0.754	0.616	0.076	0.803	0.539	0.133
	WaNet	0.949	1.000	0.175	0.109	0.989	**0.893**	**0.835**	**0.027**	0.866	0.854	**0.022**	0.691	0.602	0.273	0.757	0.818	0.034
	SIG	0.860	0.999	0.453	0.172	**0.000**	**0.857**	0.086	0.892	0.828	**0.624**	0.278	0.770	0.560	0.095	0.820	0.551	0.105
	CL	0.858	1.000	0.397	0.151	**0.000**	**0.933**	0.001	0.999	0.909	0.192	0.803	0.783	**0.679**	**0.197**	0.855	0.370	0.620
	BPPAttack	0.949	0.998	0.070	0.100	1.000	**0.926**	0.869	0.060	0.912	**0.908**	**0.017**	0.547	0.795	**0.017**	0.592	0.870	**0.011**
	Trojan	0.950	1.000	0.285	0.312	0.427	**0.867**	**0.848**	0.014	0.854	0.830	0.017	0.758	0.781	**0.006**	0.835	0.527	0.070
	LF	0.950	1.000	0.297	0.165	0.849	0.667	0.632	0.318	0.605	0.585	0.358	0.765	**0.713**	**0.000**	**0.822**	0.712	0.009
	Poison-Ink	0.952	1.000	0.272	0.148	0.414	**0.902**	**0.907**	0.021	0.889	0.892	**0.019**	0.772	0.427	0.415	0.842	0.449	0.442
	LIRA	0.952	0.991	0.312	0.249	0.171	**0.871**	0.817	0.071	0.850	0.817	0.058	0.722	0.435	**0.029**	0.761	0.397	0.212
	Average	0.937	0.999	0.281	0.202	0.451	**0.867**	0.568	0.369	0.843	**0.713**	0.233	0.734	0.642	**0.113**	0.793	0.599	0.175
GTSRB	BadNet	0.947	0.937	0.091	0.064	0.162	0.942	**0.604**	0.506	0.892	0.809	0.242	0.912	0.436	0.544	**0.951**	0.519	**0.465**
	Blend	0.952	0.997	0.055	0.055	**0.007**	0.947	0.360	0.774	0.892	**0.797**	0.237	0.958	0.657	**0.177**	**0.977**	0.613	0.221
	WaNet	0.949	0.999	0.060	0.042	0.398	**0.947**	**0.928**	0.141	0.891	0.887	0.132	0.613	0.716	0.218	0.744	0.926	**0.038**
	SIG	0.896	0.680	0.086	0.038	**0.000**	0.832	0.413	0.697	0.782	0.451	0.622	0.777	0.615	**0.026**	**0.847**	**0.626**	0.040
	CL	0.910	0.480	0.083	0.071	0.010	**0.935**	**0.921**	0.130	0.886	0.880	0.129	0.855	0.882	0.007	0.917	0.916	**0.004**
	BPPAttack	0.921	0.933	0.089	0.056	**0.036**	**0.937**	**0.931**	0.125	0.873	0.889	0.138	0.751	0.853	0.044	0.796	0.857	0.093
	Trojan	0.931	0.998	0.114	0.050	0.304	**0.936**	**0.875**	0.165	0.878	0.844	**0.123**	0.687	0.527	0.125	0.917	0.096	0.830
	LF	0.927	0.994	0.069	0.051	0.066	**0.929**	0.239	0.876	0.878	0.527	0.510	0.850	**0.777**	**0.000**	0.880	0.775	**0.000**
	Poison-Ink	0.946	0.944	0.050	0.045	0.221	0.943	**0.924**	**0.142**	0.890	0.882	0.132	0.934	0.627	0.339	**0.974**	0.453	0.524
	LIRA	0.948	0.997	0.122	0.084	0.524	**0.937**	0.830	0.246	0.888	**0.857**	**0.152**	0.922	0.804	0.515	0.957	0.447	0.555
	Average	0.933	0.876	0.082	0.056	**0.173**	**0.929**	0.703	0.380	0.875	**0.783**	0.242	0.826	0.680	**0.200**	0.906	0.623	0.277

disrupts backdoor triggers through image shrinking and padding; (2) *ZIP* [23], which applies linear transformations such as blurring to disrupt backdoor triggers and subsequently restores purified images using diffusion models; and (3) *SampDetox* [34], a recent black-box backdoor defense that leverages noise-denoising–based detoxification with diffusion models to purify poisoned images.

B. Main Results

The performance of the `Lite-BD` and its comparison with state-of-the-art black-box backdoor defenses are presented in Table I. For CIFAR-10, `Lite-BD (RE)` achieves the highest average CA and PA, alongside the lowest average ASR. Specifically for BadNets, the CA is marginally lower ($\approx 2\%$) than the top-performing ZIP; however, both ZIP and SampDetox fail completely to mitigate the attack. In our attack design, the BadNets trigger size is 6×6, which the diffusion models in ZIP and SampDetox may reconstruct, leading to defense failure. For Poison-Ink, our method achieves the second-best CA and the third-lowest ASR. On Fashion-MNIST, the `Lite-BD (RE)` yields the lowest average ASR. By considering the results of both `Lite-BD (RE)` and `Lite-BD (SW)`, we achieve the best CA in two attacks, the best

TABLE II
COMPARISON OF EXECUTION TIMES (SECONDS) FOR EACH SAMPLE. **BOLD** DENOTES THE FASTEST EXECUTION TIMES.

Dataset	Method	BadNet	Blend	WaNet	SIG	CL	Trojan	LF	Poison-Ink	BppAttack	LIRA	Average
CIFAR-10	Lite-BD (RE)	0.02	**0.02**	0.02	0.11	0.02	0.03	**0.12**	**0.04**	**0.06**	**0.13**	**0.05**
	Lite-BD (SW)	**0.02**	0.05	**0.02**	**0.04**	**0.02**	**0.03**	0.12	0.04	0.06	0.18	0.06
	SampDetox	5.12	5.15	4.79	4.96	5.02	5.55	4.91	4.83	9.42	11.96	6.17
	ZIP	1.00	1.00	1.00	1.02	1.03	1.15	0.99	0.99	2.42	6.38	1.70
Fashion-MNIST	Lite-BD (RE)	**0.02**	**0.08**	0.04	**0.12**	**0.02**	**0.03**	**0.12**	**0.12**	**0.05**	**0.14**	**0.07**
	Lite-BD (SW)	0.03	0.13	**0.02**	0.12	0.09	0.13	0.12	0.12	0.05	0.23	0.11
	SampDetox	6.34	5.89	6.37	6.09	5.51	6.91	5.60	5.85	9.59	13.11	7.13
	ZIP	1.08	1.02	1.07	1.09	1.05	1.09	0.93	1.02	2.02	3.18	1.35
GTSRB	Lite-BD (RE)	0.07	**0.08**	0.02	0.09	0.08	**0.05**	0.10	**0.04**	**0.14**	0.18	0.09
	Lite-BD (SW)	**0.07**	0.10	**0.02**	**0.03**	**0.08**	0.06	**0.10**	0.07	0.15	**0.18**	**0.08**
	SampDetox	4.25	4.61	4.42	4.41	4.28	4.95	4.15	4.01	15.02	8.97	5.91
	ZIP	1.01	1.04	1.05	0.97	1.03	1.12	0.95	0.95	4.94	4.37	1.74

TABLE III
PERFORMANCE COMPARISON CONSIDERING DIFFERENT STAGES. **S1**: STAGE 1, **S2**: STAGE 2. THIS EXPERIMENT IS DONE BY SELECTING RANDOM 1000 SAMPLES FOR EACH ATTACK.

Dataset	Attack	Baseline CA	Baseline ASR	S1 PA	S1 ASR	S2 PA	S2 ASR	S1 + S2 PA	S1 + S2 ASR
CIFAR-10	BadNet	0.950	1.000	0.919	0.040	0.040	0.982	0.921	0.017
	Blend	0.946	0.999	0.841	0.038	0.142	0.842	0.845	0.014
	WaNet	0.945	1.000	0.878	0.066	0.039	0.982	0.893	0.043
	SIG	0.866	0.999	0.044	0.953	0.632	0.032	0.484	0.226
	CL	0.865	1.000	0.868	0.000	0.069	0.951	0.868	0.000
	BPPAttack	0.945	1.000	0.907	0.024	0.035	0.987	0.907	0.012
	Trojan	0.947	1.000	0.799	0.113	0.272	0.691	0.837	0.036
	LF	0.944	0.997	0.075	0.944	0.689	0.211	0.748	0.111
	Poison-Ink	0.946	0.998	0.784	0.165	0.127	0.880	0.854	0.058

PA in three, and the lowest ASR in six scenarios. Furthermore, the gap between the best-reported metrics and our proposed method remains within 5% across four attacks. For GTSRB, Lite-BD (RE) achieves the lowest average ASR, obtaining the best CA in four attacks and the lowest ASR in six. While ZIP and SampDetox achieve higher CA and PA in some GTSRB cases, it is critical to note that their diffusion models are trained on the evaluation datasets themselves. In contrast, our method utilizes pre-trained models trained on entirely disparate datasets. Despite this disparity, our method remains competitive, achieving comparable CA and PA in most cases. Beyond defensive robustness, a primary advantage of our approach is computational efficiency. As demonstrated in Table II, our method significantly reduces execution time per sample. On average for CIFAR-10, Lite-BD (RE) and Lite-BD (SW) are **123.4**$\times$ and **102.8**$\times$ faster than Sampdetox, and **34**$\times$ and **28.33**$\times$ faster than ZIP, respectively. Similar gains are observed for Fashion-MNIST and GTSRB. This analysis confirms that Lite-BD is efficient, dataset-independent, and achieves performance comparable to or exceeding state-of-the-art black-box defenses. The purified samples from backdoor triggers for all the attacks are illustrated in Figure 3.

TABLE IV
PERFORMANCE ON MORE MODELS. THIS EXPERIMENT IS CONDUCTED USING CIFAR-10.

Model	Attack	Baseline CA	Baseline ASR	Defense PA	Defense ASR
DenseNet-121	BadNet	0.891	1.000	0.895	0.003
	Blend	0.893	0.997	0.862	0.012
	WaNet	0.891	0.991	0.882	0.025
	SIG	0.825	0.996	0.598	0.086
	LF	0.885	0.987	0.706	0.078
MobileNetV1	BadNet	0.694	1.000	0.777	0.006
	Blend	0.770	0.993	0.733	0.008
	WaNet	0.794	0.992	0.781	0.066
	SIG	0.729	0.999	0.262	0.004
	LF	0.753	0.981	0.579	0.128
ViT-Base	BadNet	0.736	1.000	0.782	0.175
	Blend	0.735	0.988	0.680	0.014
	WaNet	0.686	0.979	0.655	0.052
	SIG	0.664	0.621	0.265	0.203
	LF	0.671	0.972	0.571	0.005

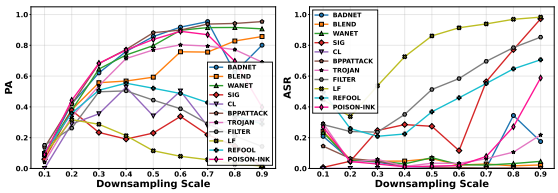

Fig. 4. Defended PA and ASR for different down-sampling scale of Stage 1. This experiment was done on CIFAR-10.

C. Impact of Different Stages of the Defense Framework

Table III illustrates the individual contributions of different stages of the `Lite-BD (RE)` across multiple backdoor attacks. From the results, we observe that most attacks can be effectively mitigated by **Stage 1**, except for **SIG** and **LF**, which require the involvement of **Stage 2**. When the trigger contains high-frequency components distributed across the spatial dimensions, as in SIG, or exhibits smooth transitions over the entire image, as in LF, band-by-band frequency filtering becomes more effective. Due to space limitations, we only present results on CIFAR-10; however, similar trends are observed on GTSRB and Fashion-MNIST.

D. Impact of Different Models on the Defense Framework

Table IV reports results on additional architectures, including DenseNet-121, MobileNetV1, and ViT-Base, under five backdoor attacks on CIFAR-10. The results demonstrate consistently strong defense performance across nearly all models and attacks, indicating that our method generalizes well across architectures. These findings confirm that our defense framework is **model-agnostic**.

E. Impact of the Downscaling Ratio on Defense Performance

Figure 4 illustrates the impact of the downscaling ratio s on prediction accuracy (PA) and attack success rate (ASR). As expected, smaller values of s lead to lower PA and lower ASR, while larger values of s yield higher PA and higher ASR. Therefore, the defender should select an intermediate value of s that balances maintaining high PA while effectively suppressing ASR. Due to space limitations, we only present results on CIFAR-10; however, similar trends are observed on GTSRB and Fashion-MNIST.

VI. Conclusion

In this paper, we propose a novel black-box backdoor defense method called `Lite-BD` that utilizes down-upscaling transformation with a pre-trained super-resolution network and a query-based band-by-band frequency filtering technique to disrupt backdoor triggers while maintaining the benign features of poisoned samples. Our method is lightweight, zero-shot, and considers both spatial and frequency transformations that are absent in state-of-the-art black-box defenses.

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019.

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

[3] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11966–11976, 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20876–20885, 2022.

[6] Yudong Gao, Honglong Chen, Peng Sun, Junjian Li, Anqing Zhang, Zhibo Wang, and Weifeng Liu. A dual stealthy backdoor: From both spatial and frequency perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1851–1859, 2024.

[7] Yudong Gao et al. A triple stealthy backdoor: Hidden in spatial, frequency, and feature domains. *IEEE Transactions on Dependable and Secure Computing*, 22(6):7746–7758, 2025.

[8] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708. IEEE, 2017.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[13] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021.

[14] Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. In *International Conference on Machine Learning*, pages 19837–19854. PMLR, 2023.

[15] Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor attack in the physical world. *arXiv preprint arXiv:2104.02361*, 2021.

[16] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020.

[17] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1833–1844, 2021.

[18] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018.

[19] Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16363–16372, 2023.

[20] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

[21] Brandon B. May, N. Joseph Tatro, Piyush Kumar, and Nathan Shnidman. Salient conditional diffusion for backdoors. In *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*, 2023.

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

[23] Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

[26] Tao Sun, Lu Pang, Weimin Lyu, Chao Chen, and Haibin Ling. Mask and restore: Blind backdoor defense at test time with masked autoencoder. *arXiv preprint arXiv:2303.15564*, 2023.

[27] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.

[28] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1905–1914, 2021.

[29] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15074–15084, 2022.

[30] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021.

[31] Zhen Xiang, Zidi Xiong, and Bo Li. Cbd: A certified backdoor detector based on local dominant probability. *arXiv preprint arXiv:2310.17498*, 2023.

[32] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.

[33] Xiaoyun Xu, Zhuoran Liu, Stefanos Koffas, Shujian Yu, and Stjepan Picek. Ban: Detecting backdoors activated by adversarial neuron noise. *arXiv preprint arXiv:2405.19928*, 2024.

[34] Yanxin Yang, Chentao Jia, Dengke Yan, Ming Hu, Tianlin Li, Xiaofei Xie, Xian Wei, and Mingsong Chen. Sampdetox: Black-box backdoor defense via perturbation-based sample detoxification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[35] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16473–16481, 2021.

[36] Jie Zhang, Dongdong Chen, Qidong Huang, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Poison ink: Robust and invisible backdoor attack. *IEEE Transactions on Image Processing*, 31:5691–5705, 2022.

[37] Shuai Zhao, Leilei Gan, Luu Anh Tuan, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.12168*, 2024.

[38] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4466–4477, 2023.