Energy-based Backdoor Defense without Task-Specific Samples and Model Retraining

Yudong Gao [1] **Honglong Chen** [1] **Peng Sun** [2] **Zhe Li** [1] **Junjian Li** [1] **Huajie Shao** [3]

Abstract

Backdoor defense is crucial to ensure the safety and robustness of machine learning models when under attack. However, most existing methods specialize in either the detection or removal of backdoors, but seldom both. While few works have addressed both, these methods rely on strong assumptions or entail significant overhead costs, such as the need of task-specific samples for detection and model retraining for removal. Hence, the key challenge is how to reduce overhead and relax unrealistic assumptions. In this work, we propose two Energy-Based BAckdoor defense methods, called EBBA and EBBA+, that can achieve both backdoored model detection and backdoor removal with low overhead. Our contributions are twofold: First, we offer theoretical analysis for our observation that a predefined target label is more likely to occur among the top results for various samples. Inspired by this, we develop an enhanced energy-based technique, called EBBA, to detect backdoored models without task-specific samples (i.e., samples from any tasks). Secondly, we theoretically analyze that after data corruption, the original clean label of a poisoned sample is more likely to be predicted as a top output by the model, a sharp contrast to clean samples. Accordingly, we extend EBBA to develop EBBA+, a new transferred energy approach to efficiently detect poisoned images and remove backdoors without model retraining. Extensive experiments on multiple benchmark datasets demonstrate the superior performance of our methods over baselines in both backdoor detection and removal. Notably, the proposed methods can effectively detect backdoored model and poisoned images as well as remove backdoors at the same time.

[1]China University of Petroleum [2]Hunan University [3]College of William & Mary. Correspondence to: Honglong Chen <chenhl@upc.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1. Whether a defense technique is model uncontrolled (MU) or can help backdoor detection in poisoned model (BD), trigger detection in poisoned image (TD), backdoor removal (BR) and not need task-specific samples (NTS) or model retraining (NMR).

Defense	MU	BD	TD	BR	NTS	NMR
MEDIC	✗	✗	✗	✓	✗	✗
ANP	✗	✗	✗	✓	✓	✗
TeCo	✓	✓	✓	✗	✗	N/A
RNP	✗	✓	✗	✓	✗	✗
Unicorn	✗	✓	✓	✗	✓	N/A
SCALE	✓	✓	✓	✗	✗	N/A
ZIP	✓	✗	✗	✓	✗	✓
EBBA (ours)	✓	✓	✓	✓	✓	✓

1. Introduction

Backdoor attacks refer to adversaries purposely manipulate either training data or model parameters to achieve accurate predictions on clean data while triggering predefined predictions on poisoned data (Gu et al., 2017). Such attacks pose a severe security threat to the safety and robustness of deep neural networks (DNNs), thereby hindering their widespread deployment in safety-critical applications such as health care and autonomous driving.

A plethora of methods have been developed to defend against backdoor attacks. In general, existing defense methods can be categorized into two types: backdoor detection and backdoor removal (Li et al., 2023b). The backdoor detection methods assess whether a model contains a backdoor (Wang et al., 2019) or if a sample contains a backdoor trigger (Guo et al., 2022; Zeng et al., 2021) while removal techniques aim to purify backdoored models by restoring their performance to that of a uncompromised model (Xu et al., 2023).

However, most existing works excel in either backdoor detection or backdoor removal. Solely detecting backdoors without removing them does not entirely address the security issue. Conversely, removing backdoors without detection lacks logical soundness. Only a few studies (Wang et al., 2019) have tackled both detection and removal, but they often rely on strong assumptions or incur high overhead, such

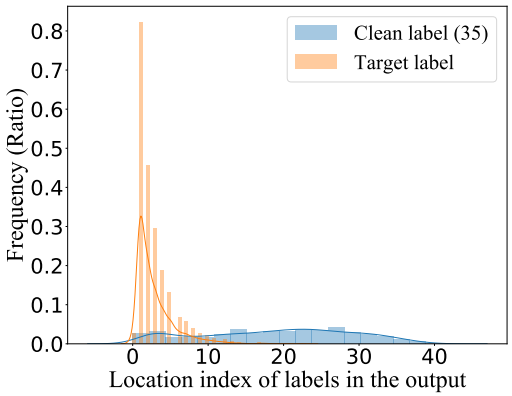

Figure 1. Location index statistical results for the target label and a clean label (label 35) on a dataset. The dataset is selectively chosen on Cifar10 and ImageNet to ensure a uniform distribution in the model outputs. The x-axis represents the position at which the labels are observed in the model's output. The output distribution of the selected dataset for the clean label is uniform. As for the target label manipulated by an attacker, it is present in the top positions in the results of the backdoored model, though it does not appear in the first position (since the image may not be fully poisoned). This indicates that despite the absence of poisoned samples in the input, the model's output still exhibits anomalies.

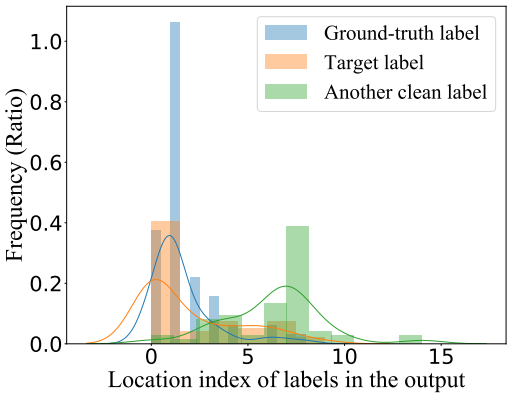

Figure 2. Location index statistical results on a set of images that are generated by a same poisoned image with 80 image corruption methods. The original location output of the poisoned image for target label, ground-truth label and another randomly selected clean label are **0** (since the image is poisoned), **6** and **7** respectively. After the poisoned image undergoes image corruption, its ground-truth label in the output has clearly shifted forward to the top results (with the portion less than 6 exceeding 90%). A small portion of the target label has shifted backward, while the randomly selected clean labels in the output remain largely unchanged.

as the need of poisoned/clean images (Liu et al., 2023) or model retraining (Li et al., 2023a) (Table 1 summarizes these limitations). In many real-world applications, accessing such resources is impractical. For instance, users typically interact with cloud-based models where model retraining is infeasible. Additionally, obtaining poisoned data for backdoor detection is usually not possible, as attackers often launch attacks at critical times, such as when autonomous vehicles encounter a poisoned road sign. These constraints severely limit the generalizability and practicality of current backdoor defenses. Therefore, the key challenge is how to mitigate these unrealistic assumptions to achieve effective and low-cost backdoor defense.

In this work, we propose two energy-based backdoor defense methods, called EBBA and EBBA+ [1], that enable the detection of backdoored models without task-specific samples and the removal of backdoors without model retraining. Here *task-specific samples* refer to clean or poisoned samples related to current tasks. Before introducing our methods, we share two insights that inspire us to develop our method. **Insight 1:** as shown in Fig. 1, we can see that the backdoored model can accurately classify the clean (even out-of-distribution) samples, but the predefined target label still has a higher likelihood of appearing among the top results. We also offer theoretical analysis to verify it in Appendix A. From the perspective of energy model (Liu et al., 2020), this suggests that the energy corresponding to the attacker's predefined target label tends to be greater than that of other labels (except for the label assigned to the sample by the poisoned model). **Insight 2:** as seen in Fig. 2, after data corruption, the original clean label of the poisoned sample is more likely to shift forward to the top outputs of the model while the target label shift backward, which is theoretically proved in Appendix C. This indicates that the energy concerning the output of a poisoned image transfers from the target label to the clean label after data corruption.

Our Contributions. Based on the above insights, we introduce two novel energy-based models for backdoor detection and removal. Inspired by insight 1, we develop an energy-based backdoor detection approach, called EBBA, that computes the energy of each label from a task-agnostic dataset (on Internet). If a label exhibits exceptionally high energy scores, we identify it as a poisoned sample. By doing this, we can detect the backdoored model without task-specific samples. According to insight 2, we extend EBBA to propose a new EBBA+ based on *transferred energy* to quantify the energy transfer phenomenon. This technique enables to accomplish poisoned images detection and backdoor removal without model retraining. Extensive experiments on three benchmark datasets demonstrate that the proposed methods outperform the baselines in both backdoor detection and backdoor removal with low overhead. Importantly, our methods can effectively detect backdoored model and poisoned images as well as remove backdoors in an all-in-one manner. What is more, our backdoor detection technique can also be applied to other classification tasks, such as speech and text classification.

[1]codes: https://github.com/ifen1/EBBA

2. Related Work

2.1. Backdoor Defenses

Generally, backdoor defenses can be grouped into two categories: backdoor detection and backdoor removal.

Backdoor Detection. (i) Backdoored model detection. The classical method for backdoored model detection involves using reverse engineering and identify the backdoor by anomaly detection, such as NC (Wang et al., 2019), Unicorn (Wang et al., 2022b) and so on. Other methods rely on poisoned samples detection to indentify the backdoored model. (ii) Poisoned samples detection. The classical approach for poisoned image detection involves using interpretable methods, such as GradCAM (Selvaraju et al., 2017), to locate the trigger's position. There are also effective methods which use output abnormality. For example, STRIP (Gao et al., 2019) superimposes various image patterns on the suspicious image to observe its output. Lower output randomness yield higher poisoning odds. Similarly, SCALE (Guo et al., 2022) superimposes the image itself and TeCo (Liu et al., 2023) corrupts the image to detect final output abnormality. Another different example is FTD (Zeng et al., 2021), which identifies poisoned samples by recognizing high-frequency noise in samples without the backdoored model. However, most of these methods (Gao et al., 2019; Guo et al., 2022; Liu et al., 2023) can only detect backdoor but not remove it.

Backdoor Removal. Backdoor removal aim to erase the backdoor from a model without compromising its classification accuracy on clean samples. One classic defense is pruning-based method. For instance, existing works, such as Fine-Pruning (Liu et al., 2018), ANP (Wu & Wang, 2021), and RNP (Li et al., 2023a), employ various methods to locate and prune backdoor neurons. Besides, ZIP (Shi et al., 2023) employs an image-reconstruction based approach to erase triggers and restore the model's performance but relies on the performance of the diffusion model. Another effective line of defense is knowledge distillation. For example, NAD (Li et al., 2021b), ARGD (Xia et al., 2022), and MEDIC (Xu et al., 2023) distill a clean model based on the loss function across multiple intermediate feature layers. (Pang et al., 2023) reconfigures a portion of the model's parameters and perform distillation using uncontaminated data. However, most backdoor removal methods require model retraining. This assumption does not hold in various scenarios like when users lack adequate computational resources and data. Table 1 summarizes the limitations of existing defenses.

Joint Backdoor Detection and Removal. In addition, very few works have studied both backdoor detection and removal. To our best knowledge, NC (Wang et al., 2019) is the first work to detect and remove backdoors using reverse engineering and neuron unlearning. However, it struggles with advanced attacks. (Li et al., 2023a) adopts an asymmetric process to reveal the backdoor neurons then prune them to achieve backdoor removal. Nevertheless, it may in turn reveal backdoor neurons in a clean model. In addition, these removal methods require model retraining. In contrast, our methods can effectively achieve backdoored model and poisoned images detection as well as backdoor removal without model retraining.

2.2. Backdoor Attack

Backdoor attack can be grouped into spatial domain backdoor and frequency domain backdoor according to the method of trigger generation.

Spatial Domain Backdoor BadNet (Gu et al., 2017) is the first to introduce the existence of backdoor attacks, placing a white or black square in the bottom right corner of an image as a trigger. Subsequently, various simple yet effective triggers were proposed, such as blended images (Chen et al., 2017), single pixels (Tran et al., 2018), and sine signals (Barni et al., 2019). However, these attacks are visually detectable, leading to recent research focus on generating visually imperceptible triggers. For example, SSBA (Li et al., 2021a) employs image steganography to create backdoors, WaNet (Nguyen & Tran, 2021) uses a distortion field to generate triggers, LIRA (Doan et al., 2021) searches for invisible triggers in high-dimensional space, and Color Backdoor utilizes intelligent algorithms to search for triggers. In fact, the invisibility of spatial domain triggers has been studied almost to the utmost.

Frequency Domain Backdoor Recently, (Zeng et al., 2021) has ventured into exploring backdoor attacks within the frequency domain. To mitigate potential high-frequency artifacts post frequency transform, a low-pass filter is utilized to create a seamless trigger. FIBA (Feng et al., 2022) crafts triggers in the frequency domain by blending the low-frequency components of two images after applying the Fourier Transform. Similarly, FTROJAN (Wang et al., 2022a) transforms the clean image using color coding methods and subsequently applies cosine transform with modifications to frequency components. However, the triggers produced in these approaches remain visible in the frequency domain. Consequently, DUBA (Gao et al., 2023a) introduces a backdoor that remains invisible in both spatial and frequency domains. Detecting backdoors from the abnormality of input samples becomes extremely challenging.

3. Preliminaries

In this section, we introduce the attacking setting and defense setting in our work as follow.

Attack Setting. Following prior works (Zhao et al., 2022;

Gao et al., 2023a), we focus on the backdoor attacks for supervised image classification, which is a widely used in various applications like face recognition and autonomous driving. Formally, the classification task needs to train a DNN model $f_\theta = f_1 \circ f_2 : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the input space, $\mathcal{Y} = \{1, 2, ..., K\}$ denotes the label space, θ indicate model parameters, f_1 represents the main structure of the model that produces the logits while f_2 is the softmax layer. The core of backdoor is to craft N_p poisoned samples $D_{\text{poison}} = \{(T(x_i), \gamma(y_i))\}_{i=1}^{N_p}$ from the training dataset $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$, where $T(\cdot)$ denotes the trigger implantation method and $\gamma(\cdot) \in \mathcal{Y}$ represents the predefined target label. When the model is trained on a considerable amount of poisoned data, it will be embedded a backdoor such that it behaves normally on the clean samples but outputs a predifined label on poisoned samples.

Defense Setting. We consider a scenario where users employ a cloud/based model. Regarding the model, the defenders can only access the model's outputs (including the last two layers) while remaining unaware of any other details, such as the model structure and loss function. Concerning the data information, the defenders can only use task-agnostic samples. They do not get any certain clean or poisoned samples when design the algorithm.

Objective. The main objective of this work is to detect backdoors in the backdoored model, identify poisoned images in a dataset, and remove backdoors to restore model performance.

4. Proposed Methods

4.1. Backdoored Model Detection via Energy Statistics

Motivation. Our insight is to shift the attention from the extremely invisible input to the output abnormality that is always neglected by attackers. As shown in Fig. 1, we can observe that even when the backdoored model is provided with clean (out-of-distribution) samples as input, the predefined target label still consistently exhibits a high probability of appearing among the top results. Below, we will introduce Lemma 1 to theoretically elucidate the observations in Fig. 1. Before that, we first introduce the two functions, f_1 and f_2, that will be used in Lemma 1.

Given an image $x^i \in \mathcal{X}$, the output Z^i (logits) of the model f_θ is:

$$f_1\left(x^i\right) = Z^i = \left\{z_1^i, z_2^i, z_3^i, ..., z_K^i\right\}, \tag{1}$$

where z_k represents the logits corresponding to the k^{th} label, and K is the the total number of labels. The softmax results of z^i is:

$$f_2(z^i) = s^i = \{s_k^i | s_k^i = \frac{e^{z_k^i}}{\sum_{k=0}^{K} e^{z_k^i}}, k \in [0, K]\}. \tag{2}$$

Lemma 1. *Suppose the poisoned model $f_\theta^p = f_1^p \circ f_2^p$, where p denotes the model has been subjected to a backdoor attack, and the attacker has a predefined target label t. Given an image x (clean or out-of-distribution) with the pseudo label k_1 (from the model f_θ^p output), $k_1 \neq t$, the model output is $f_1^p \circ f_2^p(x) = \{s_k | k \in [0, K]\}$. We have that: although s_{k_1} is greater than s_t, s_t is greater than most of s_{k_2}, where $k_2 \in [0, K]$, $k_2 \neq k_1$, and $k_2 \neq t$.*

We provide theoretical validation and additional experimental results for Lemma 1 in Appendix A and Appendix E.1, respectively.

Based on Lemma 1, we can get s_t is greater than most of s_{k_2}. We also try to explain this observation from the perspective of energy model (LeCun et al., 2006) below. When clean samples are fed into the backdoored model, the energy corresponding to the target label is significantly higher than that of other labels (excluding the label assigned to the sample by the model). Therefore, we propose utilizing energy statistics to detect the backdoored model. If a label exhibits exceptionally high energy scores, it indicates that the model is under attack. Accordingly, this specific label is more likely to be the attacker's predefined target label.

Usually, the energy of each label is determined by exponentiating the softmax results of the corresponding label. However, a direct calculation may compromise robustness in the statistical results of all samples. Imagine a well-trained model that confidently assigns the ground-truth label to a clean sample with a 99% probability (which is the case in most situations). Even if the target label receives the entire remaining 1% probability, the energy assigned to the target label will not differ significantly from that of other labels. This is because, after exponential computation, the energy of all labels becomes 1, except for the ground-truth label, which becomes extremely larger. Thus, the energy statistical results depend solely on the chosen sample distribution. Even with equal sample numbers per category, if the model is uncertain about a sample's category (e.g., assigning 80% probability) and the remaining 20% probability is unfortunately allocated to labels other than the target label, it will easily disrupt the statistical significance, rendering the statistical results essentially random. In Appendix B, we provide a more detailed explanation with examples. Therefore, it motivates us to redesign a new method to calculate the energy.

Proposed EBBA. We propose an energy-based backdoor detection model, called EBBA, that can detect the backdoored model without task-specific samples, as illustrated in Fig. 3. Specifically, we first refine the test set to achieve a uniformly distributed output, addressing concerns about skewed distributions that would lead to illogical statistical outcomes. Secondly, we set the maximum logits for each sample to 0, tackling the challenge of minor differences be-

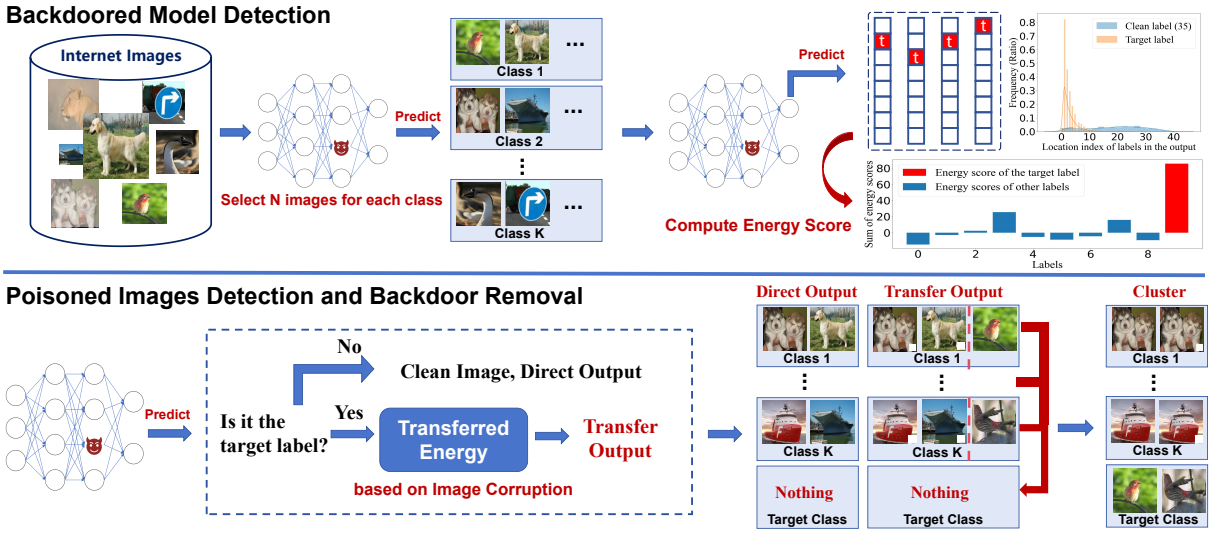

Figure 3. Overall framework of backdoored model detection, poisoned images detection, and backdoor removal. For backdoored model detection, we initially refine images obtained from the Internet to achieve a uniformly distributed output. Subsequently, we calculate the energy of each label using our improved energy method (EBBA). If the model is backdoored, the energy of the target label will be very high. For poisoned images detection and backdoor removal, we input the image into the poisoned model. If the model does not classify it into the target label (determined in the backdoored model detection phase), the image is considered clean. If the output label is the target label, we employ our proposed transferred energy method (EBBA+) to revert it to its original clean label. Thus, each clean label will have three types of images: clean images from direct output (belonging to this label), poisoned images from transfer output (belonging to this label), and clean images from transfer output (belonging to the target label). Simply employing basic binary classification methods can effectively distinguish between poisoned and clean images from the transfer output. Thus, our method can effectively achieve both trigger detection and backdoor removal.

tween the energy of labels. More specifically, since we know the model's task, we begin by collecting a large amount of data from the Internet. Then, we obtain pseudo-labels (the labels predicted by the poisoned model) for them and select a data set $\mathcal{X}^{'}$ based on these pseudo-labels to ensure an equal number of samples for each class. As mentioned before, given the image $x^i \in \mathcal{X}^{'}$, the logits of the model f_θ is Z^i and we set the maximum value in Z^i to zero (defined as MZ):

$$MZ\left(Z^i\right) = \{z_k^{i^*} = z_k^i | z_k^i = 0, k = \arg\max_k z_k^i\}. \quad (3)$$

Assume that we select N images for each class (pseudo-label), we can obtain a set of softmax results from these images after applying Eq. (3). For a clear description, we denote it by $S = \{s^1, s^2, ..., s^{N\times K}\}$. According to (Liu et al., 2020), we define the energy of label k for image x^i as:

$$P_k^i = \frac{e^{s_k^i \times T}}{T}, \quad (4)$$

where T is the temperature coefficient, s_k^i represents the value in s^i corresponding to the label k. For the purpose of convenient statistical analysis, we design a Normalize Energy as follows:

$$NP_k^i = P_k^i - \frac{e^{0\times T}}{T} = \frac{e^{s_k^i \times T} - 1}{T}, \quad (5)$$

where e^0 is the benchmark energy. Thus the energy of each label regarding dataset $\mathcal{X}^{'}$ is:

$$E = \left\{ E_k | E_k = \sum_{x^i \in \mathcal{X}^{'}} NP_k^i, k \in [0, K] \right\}. \quad (6)$$

Subsequently, a straightforward statistical method is employed to identify the label with abnormal energy. The mean (μ) and variance (σ) values for E are as follows:

$$\mu = \frac{\sum_{k=0}^{K} E_k}{K}, \sigma = \sqrt{\frac{\sum_{k=0}^{K} (E_k - \mu)}{K}}. \quad (7)$$

If there is a label $k^{'}$ that satisfies the following condition, the model is poisoned and the target label is $k^{'}$.

$$E_{k^{'}} - \mu > \lambda\sigma, \quad (8)$$

where λ is a hyper-parameter discussed in the experiments.

4.2. Poisoned Images Detection and Backdoor Removal via Transferred Energy

We extend the above EBBA to propose a new EBBA+ method for poisoned images detection and backdoor removal. In the following, we first present the motivation and then elaborate on our method.

Motivation. In prior study (Liu et al., 2023), it has demonstrated that backdoored model exhibits almost the same performance across different image corruptions for clean images but shows discrepancies for poisoned samples. However, as shown in Fig. 2, we can see that the output of poisoned samples may *not change* after image corruption. Thus, only observing the final output may result in low robustness of defense. To deal with this problem, we shift our attention from the final output to the logits. We find that even when the final output of a poisoned sample remains unchanged after corruption, its logits will be changed. We further explain how the logits change in the following Lemma 2.

Lemma 2. *Suppose the backdoored model f_θ^p has been backdoored with the target label t. Given an image x with the original ground-truth label k_1, the model output is $s = \{s_k | k \in [0, K]\}$. We apply different types of image corruptions to x to get J corruption images $D_j(x) = x^{d_j}$, where d_j represents the corrupted image generated by j^{th} corruptions method, such as Gaussian noise, raindrop effects and division by positive integers, where $j \in [1, J]$, indicating the J kinds of corruptions. The model output of x^{d_j} is $s^{d_j} = \{s_k^{d_j} | k \in [0, K]\}$. If an image x is poisoned, we have $s_t > s_t^{d_j}$ and $s_{k_1} < s_{k_1}^{d_j}$.*

We provide detailed theoretical proof of Lemma 2 in Appendix C.

We can conclude from Lemma 2 that the backdoored model shows reduced confidence in the final classification of the corrupted poisoned images. For instance, the probability of classifying it to the target label decreases from 99% to 60%, yet it is still classified to the target label. Besides, the corrupted poisoned samples are more prone to being classified back into their original ground-truth class, which means that the decrease in probability at the target label transfers to the original label.

Motivated by the above discussion, we propose a new concept of Transferred Energy (TE) to quantify this probability transfer phenomenon in Lemma 2

Proposed EBBA+. We propose a novel EBBA+ method for both poisoned images detection and backdoor removal. To this end, we propose a new concept of normalize transferred energy (NTE) as follows. For each label k, let the energy of the image x and that after image corruption (x^{d_j}) denote P_k and $P_k^{d_j}$, respectively. We define the transferred energy from image x to its corresponding corrupted image x^{d_j} of label k as:

$$\mathrm{TE}_k^j = \frac{P_k^{d_j}}{P_k} = \frac{e^{s_k^{d_j}}}{e^{s_k}} = e^{s_k^{d_j} - s_k}. \tag{9}$$

In order to conveniently count the transferred energy for each label, we normalize the TE. Since $mean(s_m^{D_j}) - mean(s_m) = 0$, the Normalized Transferred Energy (NTE) score can be defined as:

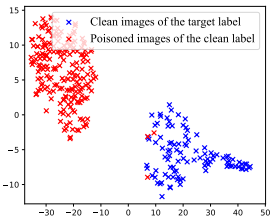

Figure 4. t-SNE on Cifar10.

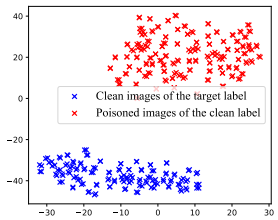

Figure 5. t-SNE on GTSRB.

$$\mathrm{NTE}_k^j = \frac{e^{s_k^{d_j}}}{e^{s_k}} - e^0 = e^{s_k^{d_j} - s_k} - 1. \tag{10}$$

Then, the NTE of all labels for image x can be expressed as:

$$\mathrm{NTE} = \{\mathrm{NTE}_k = \sum\nolimits_{j=1}^{J} \mathrm{NTE}_k^j | k \in [0, K]\}. \tag{11}$$

Next, we use NTE to purify the performance of backdoored model on poisoned images, which is a core component for poisoned images detection and backdoor removal. Concretely, if image x is poisoned, the transfer of output probability from the target to the original label leads to the NTE of the original label significantly surpassing that of other labels, while the NTE of the target label is much lower than that of other labels. Mathematically, the original label k_1 of the poisoned image x can be written as:

$$k_1 = \arg\max_k \mathrm{NTE}_k, k \in [0, K]. \tag{12}$$

While NTE can purify the backdoored model on poisoned images, it fails to identify whether image x is poisoned and achieve fully backdoor removal only by itself. The main reason is that if the image x is clean, the energy will be randomly transferred to other labels, making it hard to ensure the model performance in clean images. As a result, the NTE score cannot distinguish the clean images and poisoned images. To deal with this problem, we propose to combine it with target label identified in EBBA above. Below, we introduce how to combine these two methods to detect poisoned images and backdoor removal simultaneously.

The basic idea is that we first consider all images classified into the target class as poisoned images by the backdoored model. In this way, we only compromise the benign accuracy of the target class and locate all the poisoned images within a small range (with a few clean images of the target label). Then, we classify these images based on Eq. (12). Poisoned images will be classified into the original clean class, while clean images will be randomly classified. As a result, each clean class k will have three types of images:

poisoned/clean images belonging to class k and clean images belonging to the target class t. Based on this, we may use clustering methods to classify these images for poisoned images detection and backdoor removal simultaneously.

To achieve this, we adopt t-SNE to project the perturbed images into a feature space for clustering. We have conducted extensive experiments to demonstrate the effectiveness of our method in Appendix E.4. In particular, we present two examples in Figs. 4 and 5. It can be observed from them that our method successfully distinguishes clean samples belonging to class t from poisoned samples belonging to class k. Thus, we recover the benign accuracy of the target class and perfectly locate all the poisoned images without any clean sample. We can conclude that simple clustering methods can detect poisoned images by combining NTE score and target label.

5. Experiments

5.1. Experimental Setup

Dataset and DNN Selection. Following the settings in prior backdoor defenses (Guo et al., 2022; Shi et al., 2023), we conduct experiments on Cifar10 (Krizhevsky et al., 2009), GTSRB (Stallkamp et al., 2012) and Imagenet (Deng et al., 2009) (subset) datasets with ResNet18 (He et al., 2016). More details are presented in Appendix D.1.

Attack Methods. We evaluate our methods under five representative attacks, including BadNets (Gu et al., 2017), Blend (Chen et al., 2017), WaNet (Nguyen & Tran, 2021), FIBA (Feng et al., 2022) and DUBA (Gao et al., 2023a). The attack success rate of each attack is trained to be above 98.8% to ensure the credibility of the defense results.

Baselines. Since very few works have been developed to identify backdoor defense and removal at the same time, we compare our method with the baselines in one of the three defense types. Specifically, for backdoored model detection, we compare the proposed approach with Neural Cleanse (Wang et al., 2019), SCALE (Guo et al., 2022), Unicorn (Wang et al., 2022b), and TeCo (Liu et al., 2023). Regarding poisoned images detection, we compare it with FTD (Zeng et al., 2021), SCALE (Guo et al., 2022), and TeCo (Liu et al., 2023). In the context of backdoor removal, we compare our approach with Fine-Pruning (Liu et al., 2018), NAD (Li et al., 2021b), ANP (Wu & Wang, 2021), RNP (Li et al., 2023a), ZIP (Shi et al., 2023), and MEDIC (Xu et al., 2023).

Evaluation Metrics. For backdoored model detection, it is a binary classification problem concerning whether or not detecting the backdoor. Considering each method own its specific quantification metric, it is unfair for us to compare with them. Thus, we propose to identify whether our approach can detect the backdoor under different scenarios using binary measurement. For trigger detection in poisoned images, following TeCo (Liu et al., 2023), we adapt two metrics: Poisoned data Detection Rate (PDR) and F1 score. PDR is the ratio of true positive samples to all poisoned samples and F1 score is a more comprehensive evaluation metric. We describe the details in Appendix D.2. For backdoor removal, following (Wu & Wang, 2021; Li et al., 2023a), we adapt attack success rate (ASR) and Benign Accuracy (BA) to evaluate the defenses.

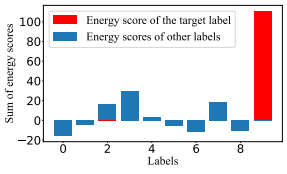

Figure 6. Energy on Cifar10.

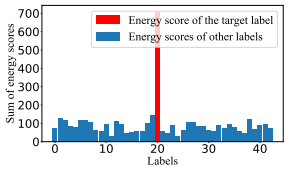

Figure 7. Energy on GTSRB.

5.2. Main Defense Results

Backdoored Model Detection. We first evaluate the proposed EBBA on backdoored model detection. The main reason is that we need to identify if the model is backdoored first so as to decide whether to perform a backdoor removal operation. As illustrated in Table 2, Neural Cleanse (NC) can not detect the backdoored model on advanced attacks. Under conditions where there are no poisoned or clean samples, only our method performs effectively, while TeCo and SCALE become entirely ineffective. Figs. 6 and 7 show two examples of the experimental results. We can see that the energy of the target label is significantly higher than other labels, showing the excellent performance of our defense. More results are presented in Appendix E.2.

Poisoned Image Detection. We also employ our EBBA+ to detect poisoned images. As we know, the backdoor behavior is triggered when the trigger and the model backdoor are present simultaneously. Table 3 illustrates the comparison results for different methods using PDR and F1. EBBA+ exhibits excellent performance in two aspects. First, it can detect all the triggers with high probability, in which the triggers are from visible to invisible and further to dual-invisible in both spatial and frequency domains. This is attributed to its property of shifting attention from the input to the output. Secondly, EBBA+ can achieve better or competitive detection performance than the baselines on three datasets. Thus, we can conclude that our proposed EBBA+ can effectively detect poisoned images.

Backdoor Removal in Poisoned Models. Lastly, we assess the performance of EBBA+ on backdoor removal. Table 4 shows the defense results of our method and baselines using BA and ASR metrics. It can be observed that EBBA+ performs the best among all the defenses in terms of BA.

Table 2. The defense results of backdoored model detection. Only the proposed EBBA can detect backdoored model in all cases.

Methods	With Poisoned/Clean Images					With Poisoned Images					Without Poisoned/Clean Images				
	BadNets	Blend	WaNet	FIBA	DUBA	BadNets	Blend	WaNet	FIBA	DUBA	BadNets	Blend	WaNet	FIBA	DUBA
NC	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗
SCALE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
Unicorn	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
TeCo	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
EBBA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 3. Defense results of poisoned images detection using PDR and F1 score (F1). The higher the better. The proposed EBBA+ outperforms all other methods in terms of PDR and F1 score.

Datasets	Backdoor Attacks	FTD PDR	FTD F1	TeCo PDR	TeCo F1	SCALE PDR	SCALE F1	EBBA+ PDR	EBBA+ F1
Cifar10	BadNets	**0.97**	0.94	0.94	0.91	**0.97**	0.93	0.96	**0.96**
	Blend	0.95	0.89	0.93	**0.95**	0.89	0.87	**0.95**	0.94
	WaNet	0.52	0.54	0.93	0.92	0.92	0.90	**0.95**	**0.94**
	FIBA	0.58	0.52	**0.96**	0.94	0.88	0.87	0.95	**0.95**
	DUBA	0.53	0.53	0.95	0.94	0.82	0.79	**0.96**	**0.95**
GTSRB	BadNets	0.96	0.93	0.92	0.88	**0.98**	0.96	**0.98**	**0.97**
	Blend	0.95	0.96	0.93	0.91	0.89	0.88	**0.97**	**0.96**
	WaNet	0.54	0.58	**0.97**	0.95	0.91	0.89	0.96	**0.95**
	FIBA	0.50	0.58	0.93	0.91	0.89	0.87	**0.95**	**0.95**
	DUBA	0.54	0.58	0.90	0.93	0.89	0.86	**0.96**	**0.95**
ImageNet	BadNets	0.90	0.93	0.91	0.92	**0.98**	**0.96**	**0.98**	0.95
	Blend	0.95	0.93	0.93	0.92	0.88	0.86	**0.98**	**0.95**
	WaNet	0.64	0.60	0.94	0.92	0.87	0.86	**0.96**	**0.94**
	FIBA	0.57	0.52	0.94	0.92	0.93	0.92	**0.96**	**0.95**
	DUBA	0.52	0.55	0.93	0.90	0.89	0.91	**0.96**	**0.95**

Besides, our method has lower or competitive ASR than the baselines under most attack scenarios. Note that the baselines except for ZIP require model retraining while ZIP try to purify samples based on the pre-trained diffusion model, which requires a large amount of clean data for pre-training. In contrast, our method does not need model retraining or pre-training.

To further show the good performance of our method, we present two examples of the NTE (Normailzed Transferred Energy) results in Figs. 8 and 9. We can observe that the ground-truth label has the highest NTE score, so our method makes the poisoned sample perform as a clean sample. Besides, since we have identified the target label in EBBA, the minimum value of BA is $\frac{K-1}{K} \times ba$, where K is the total number of labels and ba is the benign accuracy of the backdoor model. This already surpasses most defense methods. In particular, when K is large, BA can almost reach a comparable value to that in the clean model. It suggests that our method is effective in backdoor removal.

5.3. Ablation Studies and Hyper-parameter Settings

We also conduct ablation studies to explore the impact of important components on defense performance, such as the necessity of selecting Internet images. In addition, we also study the effect of important hyper-parameters on model performance, such as the threshold λ and the temperature T.

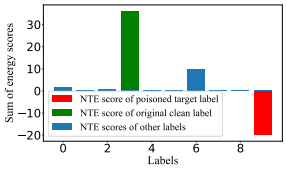

Figure 8. NTE on Cifar10.

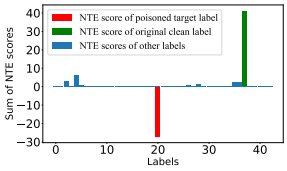

Figure 9. NTE on GTSRB.

Impact of Binary Classification. We conduct experiments on GTSRB with four clustering models, namely Hierarchical Clustering (HC), Birch, Mean Shift, and DBSCAN. The result is the average of PDR from five attack methods. As shown in Table 7, since the final result is already easily amenable to binary classification, the choice of clustering method has little impact for EBBA+.

For more details, please refer to Appendix F.

5.4. Further Exploration

We further explore the defense capability in other domains.

EBBA Against Speech and Text Classification Tasks. We find that EBBA is effective not only in image classification but also easily applicable to text and speech classifications. We conduct speech recognition experiments on ESC-50 (Piczak, 2015) and text classification experiments on THUCnews (Tnews) (Sun et al., 2016). We introduce backdoors by replacing portions of speech or characters. The defense results are illustrated in Figs. 10 and 11.

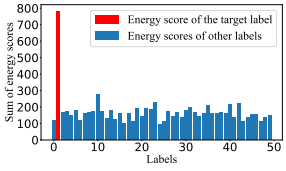

Figure 10. Energy on ESC-50.

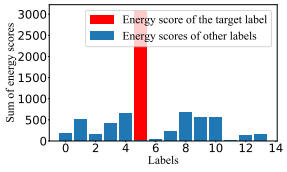

Figure 11. Energy on Tnews.

EBBA Against Multi-Label Backdoor Attacks. We test the defensive efficacy of EBBA against multi-target back-

Table 4. The defense results of backdoor removal in poisoned models. The BA of EBBA+ is best among all methods while its ASR is only comparable with them. Notably, the methods with "*" require model retraining and ZIP needs a sufficient amount of data to pre-train the diffusion model. The proposed EBBA+ does not need model retraining or additional data.

Datasets	Backdoor Attacks	No Defense		Fine-Pruning*		NAD*		ANP*		RNP*		MEDIC*		ZIP		EBBA+ (ours)	
		BA↑	ASR↓	BA↑	ASR↓	BA↑	ASR↓	BA↑	ASR↓	BA↑	ASR↓	BA↑	ASR↓	BA↑	ASR↓	BA↑	ASR↓
Cifar10	BadNets	91.22	99.36	86.79	15.76	88.96	4.72	87.08	3.42	89.72	**1.46**	88.09	3.76	88.02	5.53	**89.92**	2.92
	Blend	91.35	99.97	85.42	17.92	87.36	5.92	89.59	**1.59**	**90.06**	3.72	86.27	6.79	83.67	7.75	89.72	5.79
	WaNet	91.25	99.78	86.92	79.63	86.37	39.36	87.62	19.92	89.30	12.72	87.91	9.66	86.25	**3.96**	**89.96**	4.36
	FIBA	91.08	99.26	86.59	59.07	86.71	39.79	88.96	16.29	89.75	9.91	87.09	8.78	85.08	**2.85**	**90.02**	7.88
	DUBA	91.55	99.98	85.97	82.96	89.32	62.58	81.39	25.97	**90.95**	11.80	85.46	6.27	85.05	**4.09**	89.52	5.24
GTSRB	BadNets	99.14	99.62	94.72	7.09	92.69	2.92	95.16	2.39	97.82	**0.72**	94.58	1.99	96.92	6.19	**99.06**	**1.79**
	Blend	99.15	99.72	96.93	42.39	92.61	10.29	97.57	5.92	97.02	**2.36**	94.64	8.42	97.01	8.53	**98.92**	2.72
	WaNet	99.07	99.81	92.42	21.97	96.02	8.39	96.47	1.92	97.09	2.09	95.79	10.36	97.50	3.29	**98.79**	**1.37**
	FIBA	99.22	98.91	93.55	90.72	95.98	14.69	96.52	4.93	96.46	1.93	96.21	9.57	96.68	1.92	**98.36**	**1.47**
	DUBA	99.21	99.92	97.83	94.31	94.19	27.08	95.14	15.12	93.69	7.84	95.92	14.82	96.84	**3.33**	**98.94**	3.72
ImageNet	BadNets	88.56	99.22	79.62	2.39	85.72	9.36	85.21	11.63	85.92	6.83	86.05	2.28	87.09	7.55	**88.19**	**1.51**
	Blend	88.39	98.62	72.97	7.92	85.93	11.92	84.27	6.34	86.92	5.04	84.65	8.09	86.08	8.35	**88.36**	**1.32**
	WaNet	88.62	99.27	62.91	84.27	85.62	14.96	85.49	12.35	86.56	9.26	86.15	10.97	86.90	**3.02**	**87.94**	4.76
	FIBA	89.07	98.59	82.17	79.62	84.92	11.15	86.39	6.11	87.07	2.30	86.82	3.09	87.15	2.92	**88.09**	**1.92**
	DUBA	88.92	99.36	81.91	92.47	86.39	35.72	85.51	13.94	85.35	10.09	84.99	6.90	86.69	**2.06**	**87.91**	2.96

Table 5. Average PDR results on four clustering models.

Methods	HC	Birch	Mean Shift	DBSCAN
PDR	0.972	0.968	0.965	0.975

door attacks. Specifically, we employed the methods outlined in (Xue et al., 2020) to set up two backdoors in one model, both achieving a attack success rate of 99%. Initially, we recorded the positions where their labels appeared. As depicted in Fig. 12, the two target labels consistently appear among the top results, while clean labels tend to exhibit a uniform distribution, aligning with our expectations. The defensive outcomes using EBBA are illustrated in Fig. 13, where the energy of the two target labels significantly surpasses that of other labels. This provides evidence of EBBA's outstanding defense capabilities against multi-target attacks.

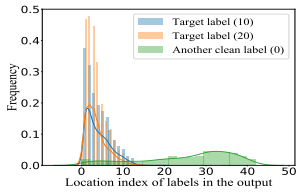

Figure 12. Location index statistical results for Two-Target attacks (target 10 and 20).

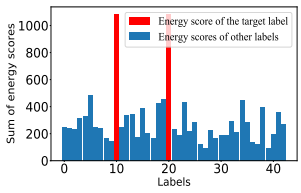

Figure 13. The results of EBBA on Two-Target attacks (target 10 and 20).

EBBA Against Adaptive Attacks. We design an adaptive attack that, during the training process, intentionally reduces the probability that a clean image belongs to the target label class by modifying the soft labels. For example, in a five-class classification task where the target label is 2 and the clean label (label 1) for clean images is one-hot encoded as [1 0 0 0 0], we modified it to [0.8 0 0.067 0.067 0.067]. This means the original clean label value is 0.8, the target label value is 0, and the rest are equally divided. We trained on the GTSRB dataset and found that this training method significantly reduces the energy value of the target label. EBBA can still detect this anomaly. Simply modifying Eq. (8) from $E_{k'} - \mu > \lambda\sigma$ to $|E_{k'} - \mu| > \lambda\sigma$ in the EBBA formula is sufficient, as the energy of the target label will exhibit an exceptionally low value under this training condition, which can still be captured by EBBA.

For more details, including more results of adaptive attacks and defense ability against clean-label attack, please refer to Appendix G.

6. Conclusion

In this paper, we developed two energy-based methods, called EBBA and EBBA+, for backdoor detection and backdoor removal. Specifically, EBBA adopted an enhanced energy statistics approach to evaluate the energy of each label from a task-agnostic dataset, enabling the detection of backdoored models without the need for clean or poisoned samples specific to the task at hand. Then we extended EBBA to propose a new EBBA+ based on transferred energy to identify poisoned images and remove backdoor simultaneously. Extensive experiments validated the superiority of our proposed methods over baselines in both backdoor detection and removal. Importantly, our approaches can provide an all-in-one defense that simultaneously detects backdoored model and poisoned images as well as removes backdoors. Furthermore, the introduced backdoor detection method can be adaptable to other classification tasks, including speech recognition and text classification.

Impact Statement

This paper presents work whose goal is to advance the field of trustworthy Machine Learning. There are many potential societal consequences of our work. Notably, our work emphasizes the enhancement of AI model security, proving advantageous for the entire community.

Acknowledgment

This work was supported in part by NSFC under Grants 61772551, 62111530052, 62102337, the Shandong Provincial Natural Science Foundation, China, under Grants ZR2023ZD32 and the Natural Science Foundation of Hunan Province of China under Grant 2023JJ40174.

References

Abbasimehr, H. and Paki, R. Improving time series forecasting using lstm and attention models. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–19, 2022.

Barni, M., Kallas, K., and Tondi, B. A new backdoor attack in cnns by training set corruption without label poisoning. In *Proceedings of the International Conference on Image Processing*, pp. 101–105, 2019.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Doan, K., Lao, Y., Zhao, W., and Li, P. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11966–11976, 2021.

Feng, Y., Ma, B., Zhang, J., Zhao, S., Xia, Y., and Tao, D. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20876–20885, 2022.

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

Gao, Y., Chen, H., Sun, P., Li, J., Zhang, A., and Wang, Z. A dual stealthy backdoor: From both spatial and frequency perspectives. *arXiv preprint arXiv:2307.10184*, 2023a.

Gao, Y., Li, Y., Zhu, L., Wu, D., Jiang, Y., and Xia, S.-T. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 139:109512, 2023b.

Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *IEEE Access*, pp. 47230–47244, 2017.

Guo, J., Li, Y., Chen, X., Guo, H., Sun, L., and Liu, C. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *Proceedings of the International Conference on Learning Representations*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

Li, Y., Li, Y., Wu, B., Li, L., He, R., and Lyu, S. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16463–16472, 2021a.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021b.

Li, Y., Lyu, X., Ma, X., Koren, N., Lyu, L., Li, B., and Jiang, Y.-G. Reconstructive neuron pruning for backdoor defense. *arXiv preprint arXiv:2305.14876*, 2023a.

Li, Y., Zhang, S., Wang, W., and Song, H. Backdoor attacks to deep learning models and countermeasures: A survey. *IEEE Open Journal of the Computer Society*, 2023b.

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

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

Liu, X., Li, M., Wang, H., Hu, S., Ye, D., Jin, H., Wu, L., and Xiao, C. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16363–16372, 2023.

Nguyen, A. and Tran, A. Wanet: Imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

Pang, L., Sun, T., Ling, H., and Chen, C. Backdoor cleansing with unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12218–12227, 2023.

Piczak, K. J. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the Annual ACM Conference on Multimedia*, pp. 1015–1018, 2015.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 618–626, 2017.

Shi, Y., Du, M., Wu, X., Guan, Z., and Liu, N. Black-box backdoor defense via zero-shot image purification. *arXiv preprint arXiv:2303.12175*, 2023.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32: 323–332, 2012.

Sun, M., Li, J., Guo, Z., Zhao, Y., Zheng, Y., Si, X., and Liu, Z. Thuctc: An efficient chinese text classifier. 2016.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. *Advances in Neural Information Processing Systems*, 31, 2018.

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., and Zhao, B. Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 707–723, 2019.

Wang, T., Yao, Y., Xu, F., An, S., Tong, H., and Wang, T. An invisible black-box backdoor attack through frequency domain. In *Proceedings of the European Conference on Computer Vision*, pp. 396–413, 2022a.

Wang, Z., Mei, K., Zhai, J., and Ma, S. Unicorn: A unified backdoor trigger inversion framework. In *Proceedings of the International Conference on Learning Representations*, 2022b.

Wu, D. and Wang, Y. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021.

Xia, J., Wang, T., Ding, J., Wei, X., and Chen, M. Eliminating backdoor triggers for deep neural networks using attention relation graph distillation. *arXiv preprint arXiv:2204.09975*, 2022.

Xu, Q., Tao, G., Honorio, J., Liu, Y., An, S., Shen, G., Cheng, S., and Zhang, X. Medic: Remove model backdoors via importance driven cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20485–20494, 2023.

Xue, M., He, C., Wang, J., and Liu, W. One-to-n & n-to-one: Two advanced backdoor attacks against deep learning models. *IEEE Transactions on Dependable and Secure Computing*, 19(3):1562–1578, 2020.

Zeng, Y., Park, W., Mao, Z. M., and Jia, R. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16473–16481, 2021.

Zhao, Z., Chen, X., Xuan, Y., Dong, Y., Wang, D., and Liang, K. Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15213–15222, 2022.

A. Proof of Lemma 1

Lemma 1. *Suppose the model $f_\theta^p = f_1^p \circ f_2^p$ has been subjected to a backdoor attack, and the attacker has predefined the target label as t. Given an image x (clean or out-of-distribution) with the pseudo label k_1 (from the model f_θ^p output), $k_1 \neq t$, the model output is $f_1^p \circ f_2^p(x) = \{s_k | k \in [0, K]\}$. We have that: although s_{k_1} is greater than s_t, s_t is greater than most of s_{k_2}, where $k_2 \in [0, K]$ and $k_2 \neq k_1 \neq t$.*

Proof. According to the NTK theory (Jacot et al., 2018), the model output of image x can be expressed as:

$$\psi(x) = \frac{\sum_{k=0}^{K} \sum_{i=0}^{n_k} \mathcal{K}(x, x_{k,i}) \cdot y_{k,i}}{\sum_{k=0}^{K} \sum_{i=0}^{n_k} \mathcal{K}(x, x_{k,i})}, \tag{13}$$

where $x_{k,i} \in \mathcal{X}$ is the training sample and $y_{k,i}$ is the corresponding one-hot label, $\psi(x) \in \mathbb{X}^K$ is an output vector with the same dimension as $y_{k,i}$ and n_k is the number of training samples for class k. Following SCALE (Guo et al., 2022), $\mathcal{K}(x, x_{k,i}) = e^{-2\gamma \|x - x_{k,i}\|^2}$,$\gamma > 0$. The model ultimately classifies image x into the class corresponding to the maximum value in the output vector.

Since $y_{k,i}$ is one-hot encoded, meaning it has a value of 1 at the ground-truth label and 0 elsewhere, the probability for each class can be computed separately. For each output value of class k, the denominator remains the same, and the probability is solely determined by the magnitude of $\sum_{i=0}^{n_k} \mathcal{K}(x, x_{k,i})$ in the numerator. That is:

$$x \in \arg\max_k \{\sum\nolimits_{i=0}^{n_k} \mathcal{K}(x, x_{k,i}) | k \in [0, K]\}. \tag{14}$$

Thus, in Lemma 1, we actually need to prove that:

$$\sum\nolimits_{i=0}^{n_t} \mathcal{K}(x, x_{t,i}) > \sum\nolimits_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i}), \tag{15}$$

where x is an image with pseudo label k_1 (from the model f_θ^p output), t is the target label, and k_2 is one of the other labels, $k_2 \neq k_1 \neq t$.

Case 1: Since the process of data poisoning is necessary, i.e., transforming an image belonging to category k_2 into a poisoned image belonging to category t, we assume that $n_t > n_{k_2}$.

$$\begin{aligned} &\sum\nolimits_{i=0}^{n_t} \mathcal{K}(x, x_{t,i}) - \sum\nolimits_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i}) \\ =&\sum\nolimits_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{t,i}) + \sum\nolimits_{i=n_{k_2}}^{n_t} \mathcal{K}(x, x_{t,i}) - \sum\nolimits_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i}) \end{aligned} \tag{16}$$

Since x is neither classified by the model into category t nor into category k_2, the value of $\sum_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{t,i})$ and $\sum_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i})$ are approximately equal. Thus,

$$\begin{aligned} &\sum\nolimits_{i=0}^{n_t} \mathcal{K}(x, x_{t,i}) - \sum\nolimits_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i}) \\ =&[\sum\nolimits_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{t,i}) - \sum\nolimits_{i=0}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i})] + \sum\nolimits_{i=n_{k_2}}^{n_t} \mathcal{K}(x, x_{t,i}) \\ \approx&0 + \sum\nolimits_{i=n_{k_2}}^{n_t} \mathcal{K}(x, x_{t,i}) \\ =&\sum\nolimits_{i=n_{k_2}}^{n_t} e^{-2\gamma \|x - x_{t,i}\|^2} \\ >&0 \end{aligned} \tag{17}$$

Case 2: Due to the low poisoning rates for many advanced attacks (1%), we consider a more general scenario where n_t equals n_{k_2}. Assume that the original clean training images of the target label t is denoted as $D_{b,t}$, which has $n_{b,t}$ clean images and the attacker adds n_p poisoned images (denoted as D_p) to it, where these n_p images are obtained by randomly sampling $n_{c,k}$ images (clean images denoted as $D_{c,k}$ and poisoned images denoted as $D_{p,k}$) from other labels, i.e., $D_p = D_{p,1} \cup D_{p,2} ... \cup D_{p,k_1} ... \cup D_{p,K}$ and $D_t = D_{b,t} \cup D_p$, denoted the training samples with label t.

$$\begin{aligned}
&\sum_{i=0}^{n_t} \mathcal{K}\left(x, x_{t,i}\right) - \sum_{i=0}^{n_{k_2}} \mathcal{K}\left(x, x_{k_2,i}\right) \\
=& \sum_{x_{t,i} \in D_t \backslash \mathrm{D}_{p,k_1}} \mathcal{K}(x, x_{t,i}) + \sum_{x_{t,i} \in \mathrm{D}_{p,k_1}} \mathcal{K}(x, x_{t,i}) - \sum_{i=0}^{n_{k_2}-n_{c,k_1}} \mathcal{K}(x, x_{k_2,i}) - \sum_{i=n_{k_2}-n_{c,k_1}}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i}).
\end{aligned} \tag{18}$$

Since x is neither classified by the model into category t nor into category k_2, the value of $\sum_{x_{t,i} \in D_t \backslash \mathrm{D}_{p,k_1}} \mathcal{K}(x, x_{t,i})$ and $\sum_{i=0}^{n_{k_2}-n_{c,k_1}} \mathcal{K}(x, x_{k_2,i})$ are approximately equal. When $x_{t,i} \in D_{p,k_1}$, $x_{t,i} = x_{c,k_1} + T$, where x_{c,k_1} is the clean image is belong to class k_1, and T is the trigger. Thus,

$$\begin{aligned}
&\sum_{i=0}^{n_t} \mathcal{K}\left(x, x_{t,i}\right) - \sum_{i=0}^{n_{k_2}} \mathcal{K}\left(x, x_{k_2,i}\right) \\
=& \sum_{x_{t,i} \in D_t \backslash \mathrm{D}_{p,k_1}} \mathcal{K}(x, x_{t,i}) - \sum_{i=0}^{n_{k_2}-n_{c,k_1}} \mathcal{K}(x, x_{k_2,i}) + \sum_{x_{t,i} \in \mathrm{D}_{p,k_1}} \mathcal{K}(x, x_{t,i}) - \sum_{i=n_{k_2}-n_{c,k_1}}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i}) \\
\approx& 0 + \sum_{x_{c,k_1} \in \mathrm{D}_{c,k_1}} \mathcal{K}(x, x_{c,k_1} + T) - \sum_{i=n_{k_2}-n_{c,k_1}}^{n_{k_2}} \mathcal{K}(x, x_{k_2,i}) \\
=& \sum_{x_{c,k_1} \in \mathrm{D}_{c,k_1}} e^{-2\gamma \left\| x - x_{c,k_1} - T \right\|^2} - \sum_{i=n_{k_2}-n_{c,k_1}}^{n_{k_2}} e^{-2\gamma \left\| x - x_{c,k_2} \right\|^2}.
\end{aligned} \tag{19}$$

Since the image x is belong to k_1 with a high probability, the value of $\sum_{x_{c,k_1} \in \mathrm{D}_{c,k_1}} e^{-2\gamma \left\| x - x_{c,k_1} \right\|^2}$ is much lager than $\sum_{i=n_{k_2}-n_{c,k_1}}^{n_{k_2}} e^{-2\gamma \left\| x - x_{c,k_2} \right\|^2}$. T is a small trigger and invisible in most case, thus,

$$\sum_{x_{c,k_1} \in \mathrm{D}_{c,k_1}} e^{-2\gamma \left\| x - x_{c,k_1} - T \right\|^2} - \sum_{i=n_{k_2}-n_{c,k_1}}^{n_{k_2}} e^{-2\gamma \left\| x - x_{c,k_2} \right\|^2} > 0. \tag{20}$$

In both case 1 and case 2, we have: $\sum_{i=0}^{n_t} \mathcal{K}\left(x, x_{t,i}\right) > \sum_{i=0}^{n_{k_2}} \mathcal{K}\left(x, x_{k_2,i}\right)$. Thus we can know that although the poisoned model is provided with clean (out-of-distribution) samples as input, the predefined target labels still consistently exhibit a high probability of appearing among the top results. Note that the above proof is conducted under the assumption of a dirty-label attack. In Appendix G.3, we provide a more general proof, demonstrating Lemma 1 under both clean-label and dirty-label settings. □

B. An Example to Further Explain Why We Redesigned a Method for Calculating Label Energy

We offer a simple and easily understandable example, assuming a model is employed for a three-class classification task (label 0, 1, 2), with the second category subjected to a backdoor attack and treated as the target class (label 1). Consider feeding three images into the model, denoted as x^1, x^2, and x^3, to achieve a uniformly distributed output. Let the logits be denoted as Z^1, Z^2, and Z^3:

$$Z^1 = [\mathbf{9}, 1, -1], \quad Z^2 = [3, \mathbf{8}, -2], \quad Z^3 = [-1, 3, \mathbf{7}]. \tag{21}$$

After applying the softmax function, we obtain the values s^1, s^2, and s^3:

$$s^1 = [\mathbf{0.99}, 0.01, 0], \quad s^2 = [0.12, \mathbf{0.88}, 0], \quad s^3 = [0.002, 0.018, \mathbf{0.98}]. \tag{22}$$

Thus the energy of label 0, 1 and 2 is:

$$E_0 = e^{0.99} + e^{0.12} + e^{0.002} = 4.82, \quad E_1 = e^{0.01} + e^{0.88} + e^{0.018} = 4.43, \quad E_2 = e^0 + e^0 + e^{0.98} = 4.62. \tag{23}$$

The result is $E_0 > E_2 > E_1$. Even though the target label consistently has the highest or second-highest probability, the ultimately calculated energy is the lowest. This occurs because when the maximum probability is exceptionally high, such as 0.99, the remaining 0.01 probability, irrespective of its assignment to any label, loses statistical significance. The final statistical outcome will be influenced by samples similar to those with uncertain results, like s^2, which is essentially a random statistical process and lacks meaningful interpretation.

Nevertheless, given our deliberate choice of a sample set with uniformly distributed outputs, theoretically, the maximum value of the output should no longer influence the statistical results. Hence, we propose setting the maximum value of the logits to 0, as follows:

$$Z^{1^*} = [0, 1, -1], \quad Z^{2^*} = [3, 0, -2], \quad Z^{3^*} = [-1, 3, 0]. \tag{24}$$

After applying the softmax function, we obtain the values s^{1^*}, s^{2^*}, and s^{3^*}:

$$s^{1^*} = [0.231, 0.66, 0.09], \quad s^{2^*} = [0.94, 0.05, 0.01], \quad s^{3^*} = [0.02, 0.94, 0.04]. \tag{25}$$

Thus the energy of label 0, 1 and 2 is:

$$E_{0^*} = e^{0.231} + e^{0.94} + e^{0.02} = 4.83, \quad E_{1^*} = e^{0.66} + e^{0.05} + e^{0.94} = 5.54, \quad E_{2^*} = e^{0.09} + e^{0.01} + e^{0.04} = 3.14. \tag{26}$$

The result is $E_{1^*} > E_{0^*} > E_{2^*}$. When the number of images is large, the difference will become more pronounced.

C. Proof of Lemma 2

Lemma 2. *Suppose the model f_θ^p has been backdoored with the target label t. Given an image x with the original ground-truth label k_1, the model output is $s = \{s_k | k \in [0, K]\}$. We apply different types of image corruptions to x to get J corruption images $D_j(x) = x^{d_j}$, such as gaussian noise, raindrop effects and division by positive integers, where $j \in [1, J]$, indicating the J ways of corruptions. The model output of x^{d_j} is $s^{d_j} = \{s_k^{d_j} | k \in [0, K]\}$. If the image x is poisoned, we have that: $s_t > s_t^{d_j}$ and $s_{k_1} < s_{k_1}^{d_j}$.*

Proof. We demonstrate that one of the data augmentation methods satisfies the Lemma 2. In fact, as long as one data augmentation method proves effective, the subsequent transfer energy will be valid. We choose to prove the effectiveness of division by positive integers. Other augmentation methods can be demonstrated in the similar way.

Following SCALE (Guo et al., 2022), we assume that the model has only two classes with labels 0 and 1. Set label 1 as the target label and 0 as the clean label. Given a clean image x with the original ground-truth label 0, $x^p = x + T$ is the poisoned image which belongs to calss 1. We get the corrupted image $x^d = x^p/n$, where n is a positive integer greater than 1 and assume that there are N_b clean samples (denoted as D_b) and N_p poisoned samples (denoted as D_p), where clean samples can be divide into two subsets, i.e., $D_{b,0}$ and $D_{b,1}$ belong to class 0 and 1, respectively. We can rewrite the NTK expression, Eq. (13), as:

$$\begin{aligned}\psi(x) &= \frac{\sum\limits_{X \in D_{b,0}} \mathcal{K}(x, X) \cdot 0 + \sum\limits_{X \in D_{b,1}} \mathcal{K}(x, X) \cdot 1 + \sum\limits_{X \in D_p} \mathcal{K}(x, X) \cdot 1}{\sum\limits_{X \in D_{b,0}} \mathcal{K}(x, X) + \sum\limits_{X \in D_{b,1}} \mathcal{K}(x, X) + \sum\limits_{X \in D_p} \mathcal{K}(x, X)} \\ &= \frac{\sum\limits_{X \in D_{b,1} \cup D_p} \mathcal{K}(x, X)}{\sum\limits_{X \in D_b \cup D_p} \mathcal{K}(x, X)}.\end{aligned} \tag{27}$$

In fact, we only need to prove that:

$$\psi(x^p/n) < \psi(x^p) \Rightarrow \frac{\sum\limits_{X \in D_{b,1} \cup D_p} \mathcal{K}(x^p/n, X)}{\sum\limits_{X \in D_b \cup D_p} \mathcal{K}(x^p/n, X)} < \frac{\sum\limits_{X \in D_{b,1} \cup D_p} \mathcal{K}(x^p, X)}{\sum\limits_{X \in D_b \cup D_p} \mathcal{K}(x^p, X)}. \tag{28}$$

We first establish the proof for $\mathcal{K}(x^p/n, X) < \mathcal{K}(x^p, X)$, where $X \in D_{b,1} \cup D_p$, and then proceed to prove Eq. (28). Thus we need to prove:

$$\begin{aligned} & e^{-2\gamma \|x^p/n - X\|^2} < e^{-2\gamma \|x^p - X\|^2}, \gamma > 0 \\ \Rightarrow & \left\| \frac{x + T}{n} - X \right\|^2 > \|x + T - X\|^2. \end{aligned} \tag{29}$$

Please note that x, T, and X are matrices. For each element in these three matrices, denoted as a, t and b respectively, if

they all satisfy Eq. (29), then it is guaranteed to hold true. Thus we prove the following expression:

$$
\begin{aligned}
&\left(\frac{a+t}{n}-b\right)^2>(a+t-b)^2\\
\Rightarrow&\frac{(a+t)^2}{n^2}-\frac{2b\,(a+t)}{n}+b^2>(a+t)^2-2b\,(a+t)+b^2\\
\Rightarrow&\frac{(a+t)^2}{n^2}-\frac{2b\,(a+t)}{n}>(a+t)^2-2b\,(a+t)\\
\Rightarrow&\frac{(a+t)}{n^2}-\frac{2b}{n}>a+t-2b\\
\Rightarrow&\frac{(a+t)}{n^2}-(a+t)>\frac{2b}{n}-2b\\
\Rightarrow&(1-n^2)(a+t)>2n(1-n)b\\
\Rightarrow&(1+n)(a+t)<2nb\\
\Rightarrow&t<\frac{2nb}{1+n}-a.
\end{aligned}
\tag{30}
$$

Since n is a positive integer greater than 1, $\frac{2n}{1+n}\in[\frac{4}{3},2)$. We need to prove: $t<\frac{4}{3}b-a$. Note that we only need to demonstrate the validity of Eq. (29), which essentially calculates the sum of squared Euclidean distances. Therefore, it is sufficient for the majority of elements t in the trigger matrix T to satisfy $t<\frac{4}{3}b-a$.

If $X\in D_{b_1}$, a and b are two pixels of two images from class 0 and class 1, $(b-a)$ is large enough in most cases and $(\frac{4}{3}b-a)$ is larger.

If $X\in D_p$, a is the pixel of the clean image from class 0 and b is the pixel of the poisoned image made by class 0, $(b-a)$ is equal to t in most cases and $(\frac{4}{3}b-a)$ is clearly larger than t. Thus Eq. (29) clearly holds and has a greater probability to hold as n increases. Now we have that: $\mathcal{K}(x^p/n,X)<\mathcal{K}(x^p,X)$, where $X\in D_{b,1}\cup D_p$. Thus:

$$
\begin{aligned}
&\frac{\sum\limits_{X\in D_{b,1}\cup D_p}\mathcal{K}(x^p/n,X)}{\sum\limits_{X\in D_b\cup D_p}\mathcal{K}(x^p/n,X)}\\
=&\frac{\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p/n,X)\right]\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}{\left[\sum\limits_{X\in D_{b,0}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p/n,X)\right]\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}\\
<&\frac{\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p/n,X)\right]\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}{\left[\sum\limits_{X\in D_{b,0}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p/n,X)\right]}\\
=&\frac{\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}{\left[\sum\limits_{X\in D_{b,0}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}.
\end{aligned}
\tag{31}
$$

Since $\mathcal{K}(x^p/n,X)<\mathcal{K}(x^p,X)$ $(X\in D_{b,1})$, $\mathcal{K}(x^p/n,X)>\mathcal{K}(x^p,X)$ $(X\in D_{b,0})$. Otherwise x^p will not belong to any

category. Thus,

$$
\begin{aligned}
&\frac{\sum\limits_{X\in D_{b,1}\cup D_p}\mathcal{K}(x^p/n,X)}{\sum\limits_{X\in D_b\cup D_p}\mathcal{K}(x^p/n,X)}\\
=&\frac{\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}{\left[\sum\limits_{X\in D_{b,0}}\mathcal{K}(x^p/n,X)+\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}\\
<&\frac{\left[\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}{\left[\sum\limits_{X\in D_{b,0}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_{b,1}}\mathcal{K}(x^p,X)+\sum\limits_{X\in D_p}\mathcal{K}(x^p,X)\right]}\\
=&\frac{\sum\limits_{X\in D_{b,1}\cup D_p}\mathcal{K}(x^p,X)}{\sum\limits_{X\in D_b\cup D_p}\mathcal{K}(x^p,X)}.
\end{aligned}
\tag{32}
$$

□

D. Details of Experiment Settings

D.1. Datasets

To comprehensively assess the performance across various tasks, we conducted experiments utilizing three datasets: Cifar10, which is a dataset for object classification encompassing a diverse range of items including horses and aircraft. Gtsrb, a dataset specifically tailored for traffic signal recognition. ImageNet, a renowned dataset for object classification, containing an extensive array of objects. Due to its vast size, we opted to work with a subset of ImageNet. The specific dataset details are outlined in Table 6.

Table 6. Dataset information.

Datasets	Training/Testing Size	Lables Size	Image Size
Cifar10	50000/10000	10	32×32×3
Gtsrb	39209/12603	43	64×64×3
ImageNet	48000/12000	100	224×224×3

D.2. Evaluation Metrics

For trigger detection in poisoned images, we adapt Poisoned data Detection Rate (PDR) and F1 score as metrics. The detail expressions are as follows:

$$PDR=\frac{TP}{TP+FN}, \tag{33}$$

where TP means True Positive samples and FN means False Negative samples. Thus PDR represents the proportion of detected poisoned samples out of all poisoned samples.

$$F1\ score = \frac{2 \times Precision \times PDR}{Precision + PDR}, \tag{34}$$

where $Precision=\frac{TP}{TP+FP}$ and FP stands for the False Positive samples. Thus F1 score is a metric that balances precision and PDR (also named recall), providing a single value that reflects a model's overall performance in binary classification tasks.

D.3. Implantation Details

In EBBA, when testing a backdoored model trained on one of the three datasets (Cifar10, GTSRB and ImageNet), the other two datasets with some task-agnostic images from Internet are used as out-of-distribution (OOD) datasets for evaluation. The threshold λ is set to 3. The temperature T is set to 2 and the Birch is chosen as the final clustering model.

To effectively test the proposed EBBA, the backdoor attacks are set as follows: For BadNets, a black block is injected into the image; For Blend, the blend ratio is set as 0.2, and the default "hello kitty" pattern is adopted; For WaNet, we set the uniform grid of size to 6; For FIBA, we typically enhance the embedding strength and ratio to ensure the ASR. For DUBA, we adopt the default setting; We utilize the stochastic gradient descent (SGD) optimizer to train the backdoored model over a span of 200 epochs. The learning rate is established at 0.01, accompanied by a decay factor of 0.1 and decay intervals occurring at epochs 50, 100, and 150. A batch size of 64 is employed for the training process.

E. More Detailed Experimental Results

E.1. More Results of Location Index statistic

More results of motivation for backdoored model detection. As shown in Fig. 14 to Fig. 21, we present additional experimental results on the positioning of target labels and a specific clean label in the final outcome after inputting datadet into the poisoned model. The dataset is selectively chosen to ensure a uniform distribution in the model outputs. The results indicate that the target label consistently appears among the top positions, while the position of the clean label tends to be more uniformly distributed.

More results of motivation for poisoned images detection and backdoor removal. From Fig. 22 to Fig. 25, each figure presents the statistical results of the location index on a set of images generated by the same poisoned image using 80 image corruption methods. These results indicate that following image corruption of the poisoned image, the ground-truth label in the output has notably shifted forward. A minor portion of the target label has shifted backward, whereas the randomly selected clean labels in the output largely remain unchanged.

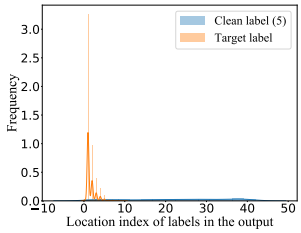

Figure 14. Label 5 of BadNet.

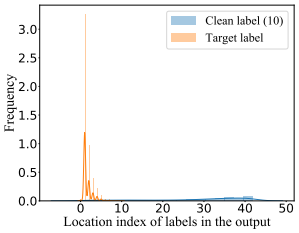

Figure 15. Label 10 of BadNet.

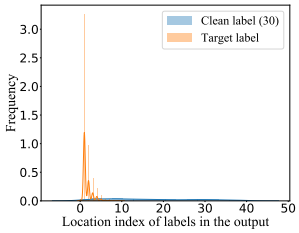

Figure 16. Label 30 of BadNet.

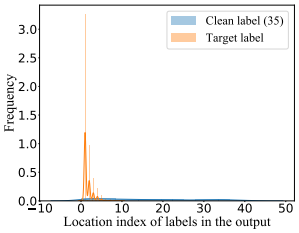

Figure 17. Label 35 of BadNet.

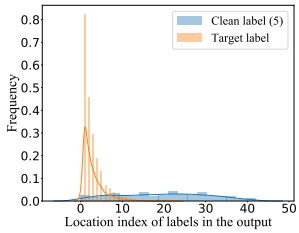

Figure 18. Label 5 of WaNet.

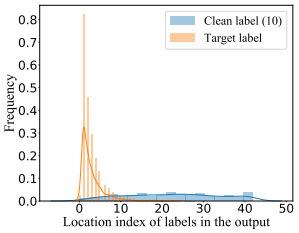

Figure 19. Label 10 of WaNet.

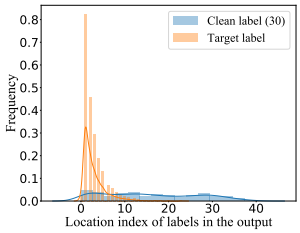

Figure 20. Label 30 of WaNet.

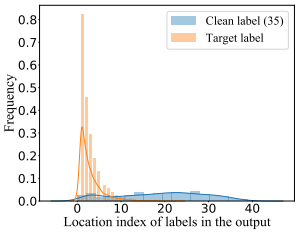

Figure 21. Label 35 of WaNet.

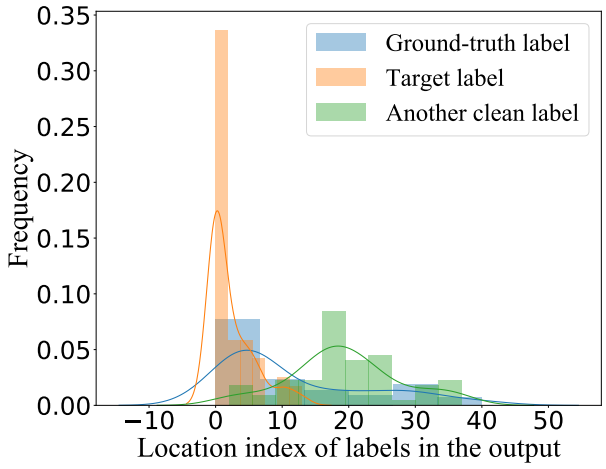

Figure 22. The original location of the poisoned image for target label, ground-truth label and clean label are 0, 25 and 20.

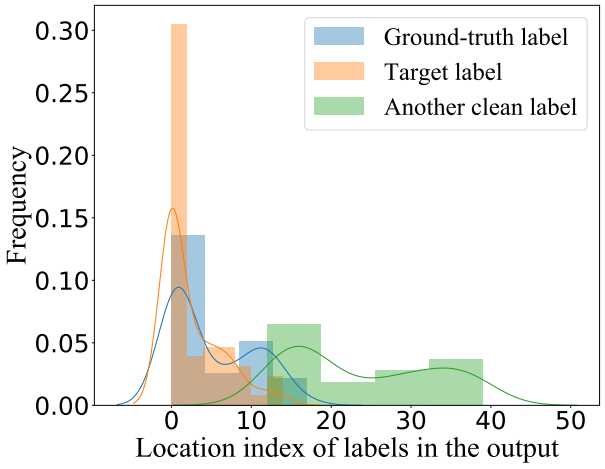

Figure 23. The original location of the poisoned image for target label, ground-truth label and clean label are 0, 10 and 22.

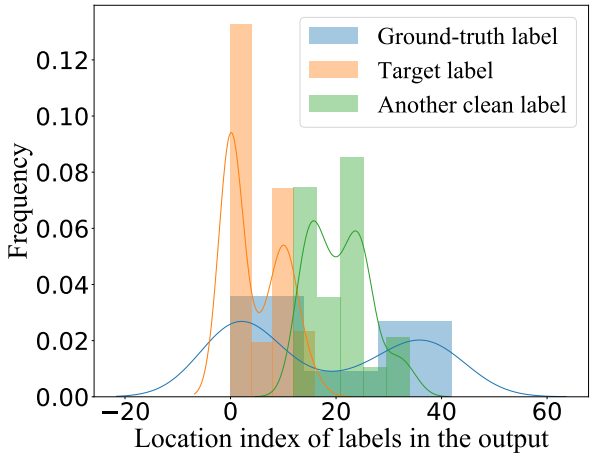

Figure 24. The original location of the poisoned image for target label, ground-truth label and clean label are 0, 38 and 20.

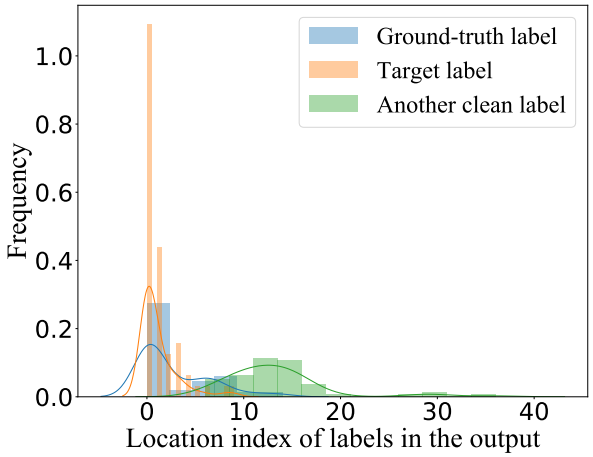

Figure 25. The original location of the poisoned image for target label, ground-truth label and clean label are 0, 7 and 12.

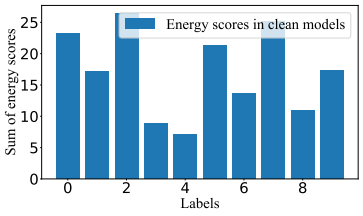

Figure 26. Clean Model of Cifar10.

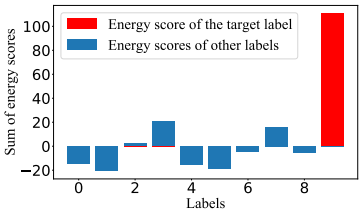

Figure 27. BadNets on Cifar10.

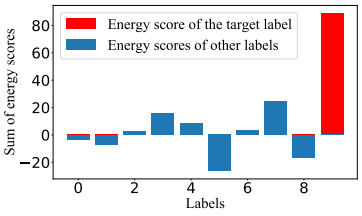

Figure 28. Blend on Cifar10.

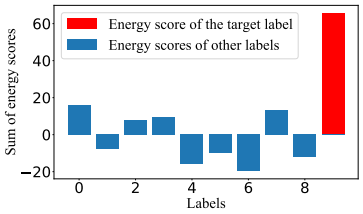

Figure 29. WaNet on Cifar10.

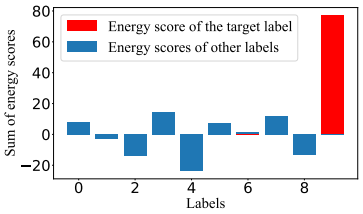

Figure 30. FIBA on Cifar10.

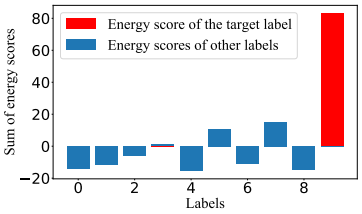

Figure 31. DUBA on Cifar10.

E.2. More Results of Energy Statistic for Backdoor Detection in Poisoned Models

As shown in Fig. 26 to Fig. 43, we demonstrated the effectiveness of EBBA in detecting backdoors across different datasets and attack methods. In comparison to results from a clean model, the energy of target labels in the attacked model is noticeably anomalous. This strongly highlights the efficacy of EBBA in detecting backdoor models.

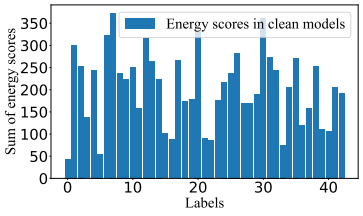

Figure 32. Clean on GTSRB.

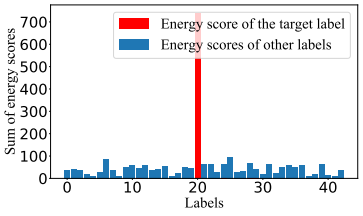

Figure 33. BadNets on GTSRB.

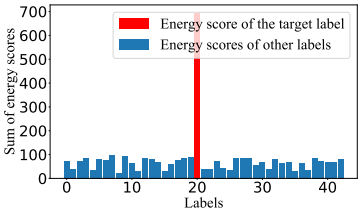

Figure 34. Blend on GTSRB.

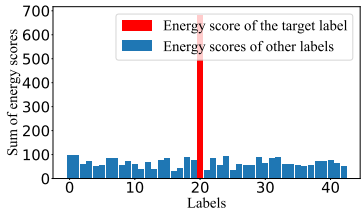

Figure 35. WaNet on GTSRB.

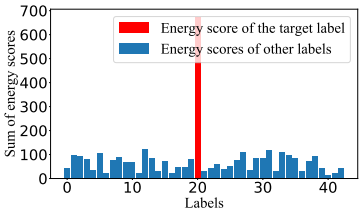

Figure 36. FIBA on GTSRB.

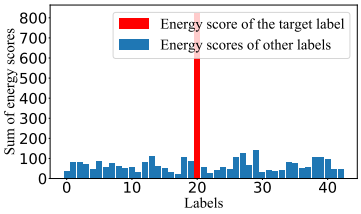

Figure 37. DUBA on GTSRB.

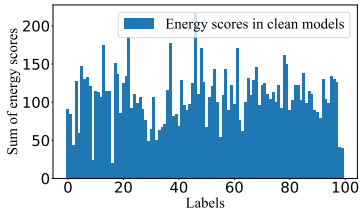

Figure 38. Clean Model of ImageNet.

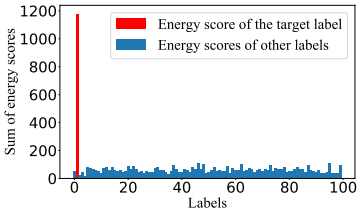

Figure 39. BadNets on ImageNet.

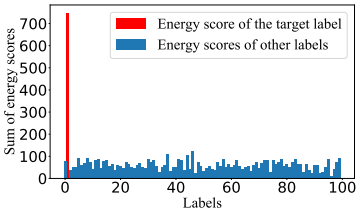

Figure 40. Blend on ImageNet.

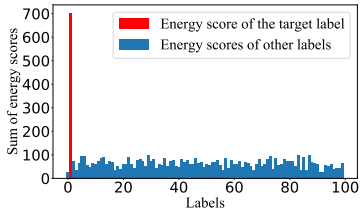

Figure 41. WaNet on ImageNet.

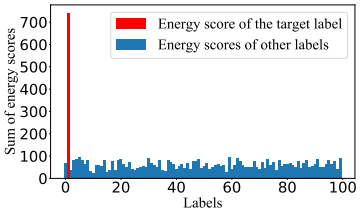

Figure 42. FIBA on ImageNet.

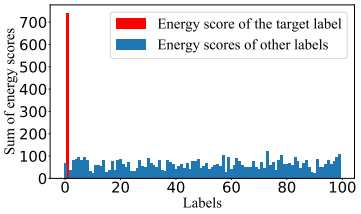

Figure 43. DUBA on ImageNet.

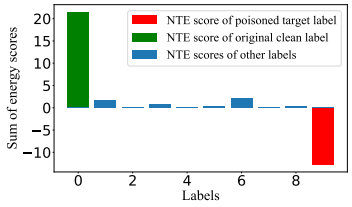

Figure 44. Label 0 of Cifar10.

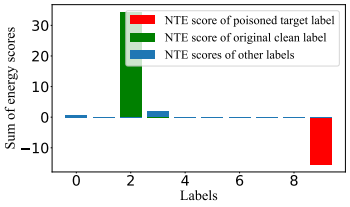

Figure 45. Label 2 of Cifar10.

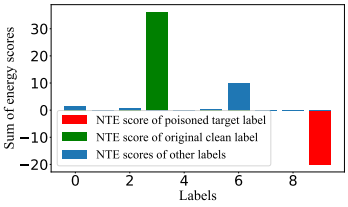

Figure 46. Label 3 of Cifar10.

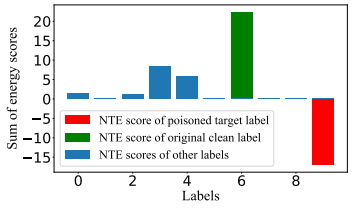

Figure 47. Label 6 of Cifar10.

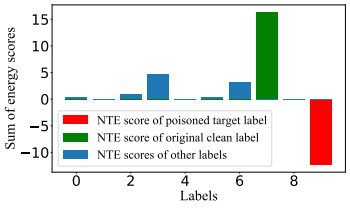

Figure 48. Label 7 of Cifar10.

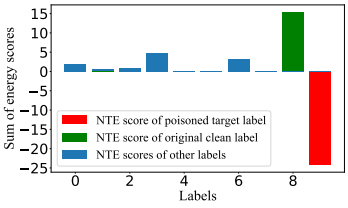

Figure 49. Label 8 of Cifar10.

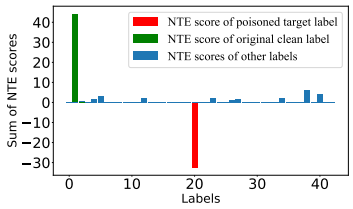

Figure 50. Label 1 of GTSRB.

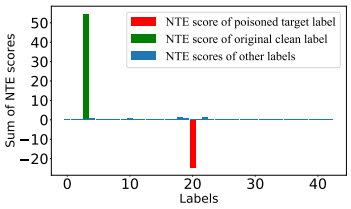

Figure 51. Label 3 of GTSRB.

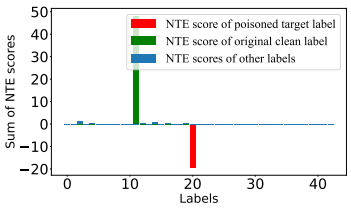

Figure 52. Label 11 of GTSRB.

E.3. More Results of the NTE Scores for Trigger Detection and Backdoor Removal

As illustrated in Fig. 44 to Fig. 61, we present additional experimental results in EBBA, focusing on the NTE scores for various samples. When an image is poisoned, its NTE score is calculated, resulting in a significant decrease in the score for the poisoned target label. Simultaneously, the score for the original clean label becomes substantially higher, while the scores for the remaining labels hover around 0.

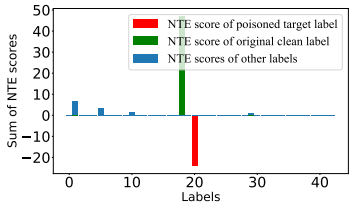

Figure 53. Label 18 of GTSRB.

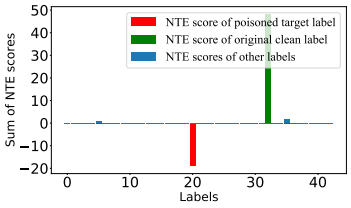

Figure 54. Label 32 of GTSRB.

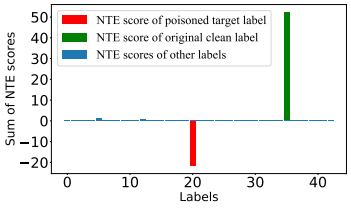

Figure 55. Label 35 of GTSRB.

E.4. More Results of the Final t-SNE for Trigger Detection and Backdoor Removal

As shown in Fig. 62 to Fig. 77, we present additional t-SNE results. Specifically, we showcased t-SNE results for poisoned data belonging to a certain clean label and clean data belonging to the target label, given that we are dealing with a binary classification task. In Cifar10, where the poisoned label is 9, we display t-SNE results for poisoned data originally belonging to labels 0 - 7 and clean data belonging to label 9. In GTSRB, the poisoned label is 20. The results demonstrate that t-SNE effectively separates the two classes of images. This underscores the high detection rate of poisoned samples and the strong performance in removing backdoors achieved by EBBA.

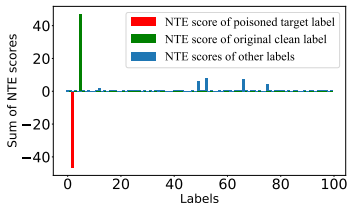

Figure 56. Label 5 of ImageNet.

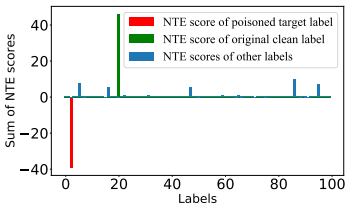

Figure 57. Label 20 of ImageNet.

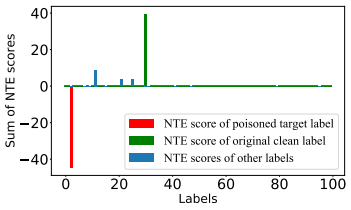

Figure 58. Label 30 of ImageNet.

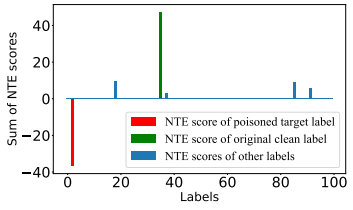

Figure 59. Label 35 of ImageNet.

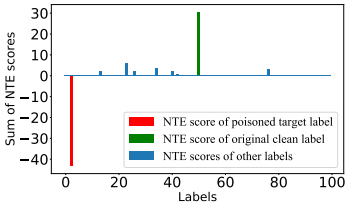

Figure 60. Label 50 of ImageNet.

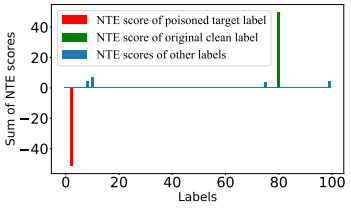

Figure 61. Label 80 of ImageNet.

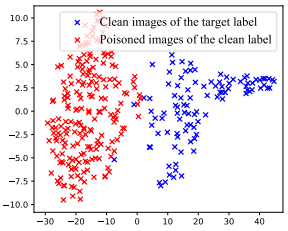

Figure 62. Label 0 of Cifar10.

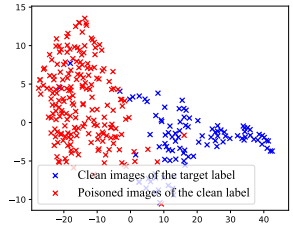

Figure 63. Label 1 of Cifar10.

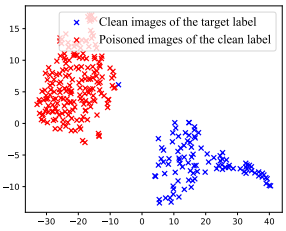

Figure 64. Label 2 of Cifar10.

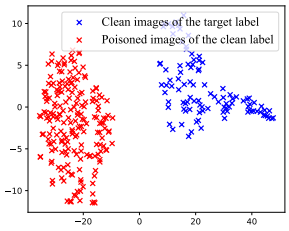

Figure 65. Label 3 of Cifar10.

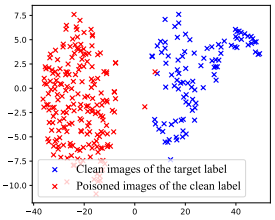

Figure 66. Label 4 of Cifar10.

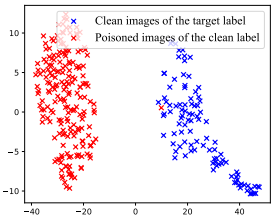

Figure 67. Label 5 of Cifar10.

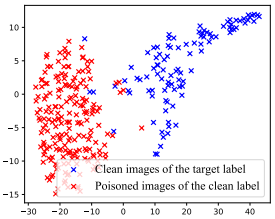

Figure 68. Label 6 of Cifar10.

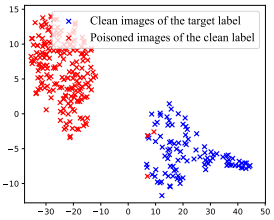

Figure 69. Label 7 of Cifar10.

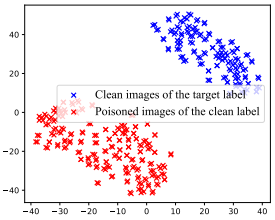

Figure 70. Label 0 of GTSRB.

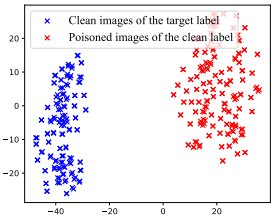

Figure 71. Label 5 of GTSRB.

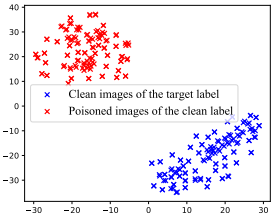

Figure 72. Label 10 of GTSRB.

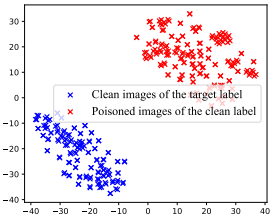

Figure 73. Label 15 of GTSRB.

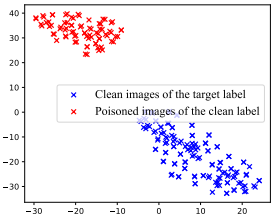

Figure 74. Label 25 of GTSRB.

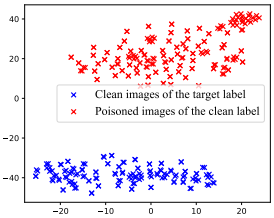

Figure 75. Label 30 of GTSRB.

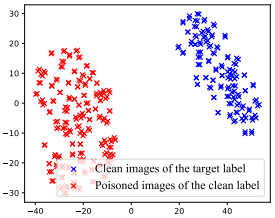

Figure 76. Label 35 of GTSRB.

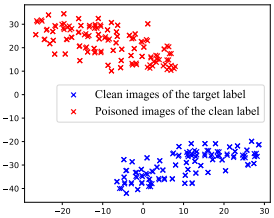

Figure 77. Label 40 of GTSRB.

F. More details of Ablation Study

Necessity of Selecting Internet Images. In backdoored model detection, the first step is to select a dataset that results in a uniformly distributed output from the model. We remove this component to emphasize the importance of this configuration. As shown in Figures 78 and 79, although the energy of the target label remains the highest, the energy of several clean labels is also close to that of the target label. As a result, multiple false positive labels are ultimately detected. This is due to the dataset selected being biased towards certain classes, leading to a final result that is skewed towards specific labels. This highlights the importance of carefully choosing the dataset.

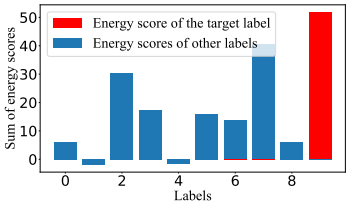

Figure 78. Energy on Cifar10 (without selecting images).

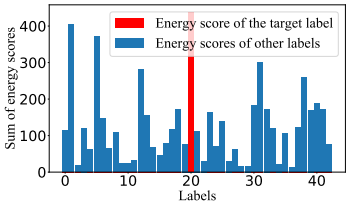

Figure 79. Energy on GTSRB (without selecting images).

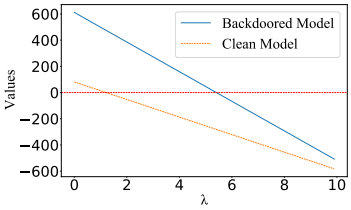

Figure 80. Impact of the threshold λ on $E_{k'} - \mu - \lambda\sigma$.

Impact of the threshold λ. We conduct experiments on the Impact of threshold λ in EBBA for backdoored model detection. We analyse the impact of λ on the value of $E_{k'} - \mu - \lambda\sigma$. If this value is lower than 0, the model is considered as backdoored. We conducted comparative experiments on poisoned and clean models. The magnitude of λ should be chosen such that the value for the clean model is less than 0, while the value for the poisoned model is greater than 0. As shown in Figure 80, to meet the above conditions, the value of λ should be approximately between 1.8 and 5.7. Therefore, we ultimately chose lambda to be 3.

Impact of Binary Classification. We conduct experiments on GTSRB with four clustering models, namely Hierarchical Clustering (HC), Birch, Mean Shift, and DBSCAN. The result is the average of PDR from five attack methods. As shown in Table 7, since the final result is already easily amenable to binary classification, the choice of clustering method has little impact for EBBA+.

Table 7. Average PDR results on four clustering models.

Methods	HC	Birch	Mean Shift	DBSCAN
PDR	0.972	0.968	0.965	0.975

Impact of the Temperature T. We conduct experiments on the impact of temperature T on the energy statistical results for backdoored model detection. As shown in Figure 81, we still analyse the impact of T on $E_{k'} - \mu - 3\sigma$. When T increases, this value also becomes larger, demonstrating that the detection performance improves. But this does not mean that a larger T is always better, as a very large T may cause the detection method to focus only on the second largest value, rendering the significance of setting the largest value to 0 meaningless. Thus we set T to be 2.

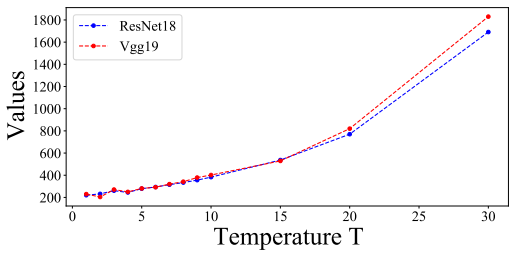

Figure 81. Impact of Temperature T on $E_{k'} - \mu - 3\sigma$.

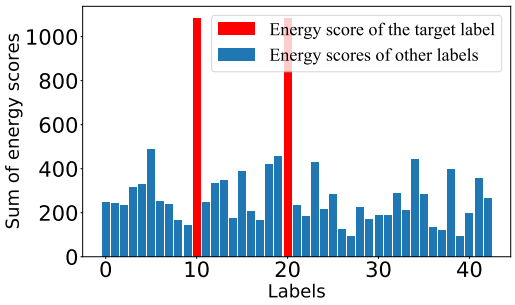

Figure 82. Energy of Two-Target Backdoors.

G. Further Exploration

G.1. More Details about EBBA Against Speech and Text Classification Tasks

We conduct defensive testing experiments for EBBA on the speech dataset ESC-50 (Piczak, 2015) and the text dataset THUCnews (Sun et al., 2016).

Speech ESC-50. The ESC-50 dataset is a labeled collection comprising 2000 environmental audio recordings intended for evaluating methods in environmental sound classification. The dataset consists of 5-second-long recordings categorized into 50 semantical classes, each containing 40 examples, such as dog, pig, and cat. We divide it into training and test sets in a ratio of 4:1. The model PANN-CNN14 (Kong et al., 2020) is employed to train the backdoor. The backdoor trigger is generated through a simple substitution similar to the BadNets method, wherein a portion of the audio in the samples is replaced and used as the trigger. The final accuracy on the clean test set is 93%, and the ASR on the poisoned test set is 99.69%. This aligns with our testing requirements.

Text THUCnews. THUCnews is generated by filtering historical data from the Sina News RSS subscription channel between 2005 and 2011. It consists of 14 categories, such as finance, lottery, and real estate. The dataset includes 752,471 examples in the training set and 83,599 examples in the test set. The model LSTM with Muti-Head Attention module (Abbasimehr & Paki, 2022) is employed to train the backdoor. The backdoor trigger is generated by replacing a specific character in the text, similar to the BadNets method. The final classification accuracy on clean samples is 97.01%, and the ASR on poisoned samples is 100%.

G.2. Defense Capability Against Multi-Label Backdoor Attacks.

We test the defensive efficacy of EBBA against multi-target backdoor attacks. Specifically, we employed the methods outlined in (Xue et al., 2020) to set up two backdoors in one model, both achieving a attack success rate of 99%. Initially, we recorded the positions where their labels appeared. As depicted in Figs. 83 to 86, the two target labels consistently appear among the top results, while clean labels tend to exhibit a uniform distribution, aligning with our expectations. The defensive outcomes using EBBA are illustrated in Fig. 82, where the energy of the two target labels significantly surpasses that of other labels. This provides evidence of EBBA's outstanding defense capabilities against multi-target attacks.

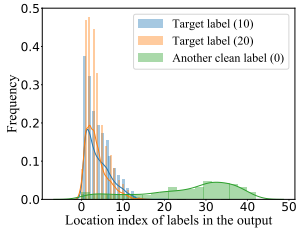

Figure 83. Label 10 of GTSRB.

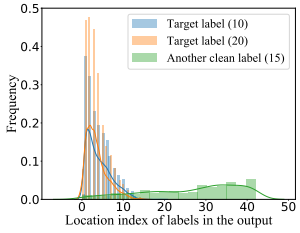

Figure 84. Label 15 of GTSRB.

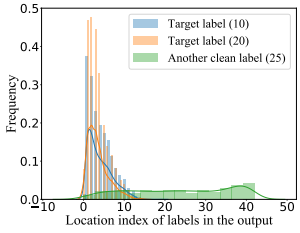

Figure 85. Label 25 of GTSRB.

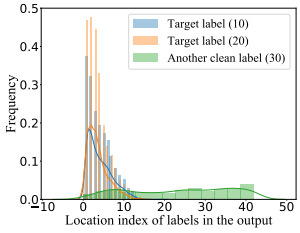

Figure 86. Label 30 of GTSRB.

G.3. The Defense Capability of EBBA Against Clean Label Backdoor

In previous experiments and demonstrations, we show EBBA's robust detection and removal capabilities against various advanced backdoors. In this section, we will further validate EBBA's ability to detect clean label backdoors. Specifically, this involves embedding a trigger into an image without modifying its label, thus creating a more invisible backdoor.

In the proof of Lemma 1 (appendix A), it is important to highlight that our assumption revolves around the presence of a dirty-label backdoor, implying a modification of labels. Therefore, we first supplement the proof to demonstrate EBBA's detection capability for clean label backdoors. In fact, we still need to prove Eq. (15) in the appendex A.

Proof. We assume that n_t is equal to n_{k_2}. Despite being a clean label backdoor attack, its behavior during the testing phase remains identical to that of dirty-label backdoor attacks. That is, the image x is originally assigned to category k_1 by the model, but with the addition of the trigger T, $x+T$ belongs to the target label t. We have:

$$\begin{aligned}
&\sum_{i=0}^{n_t} \mathcal{K}\left(x, x_{t,i}\right) - \sum_{i=0}^{n_{k_2}} \mathcal{K}\left(x, x_{k_2,i}\right) \\
=&\sum_{i=0}^{n_t} \mathcal{K}\left(x+T-T, x_{t,i}\right) - \sum_{i=0}^{n_{k_2}} \mathcal{K}\left(x, x_{k_2,i}\right) \\
=&\sum_{i=0}^{n_t} \left(e^{-2\gamma\|x+T-x_{t,i}-T\|^2} - e^{-2\gamma\left\|x-x_{k_2,i}\right\|^2}\right) \\
>&\sum_{i=0}^{n_t} \left(e^{-2\gamma\|x+T-x_{t,i}\|^2-2\gamma\|T\|^2} - e^{-2\gamma\left\|x-x_{k_2,i}\right\|^2}\right)
\end{aligned} \tag{35}$$

Since $x+T$ belongs to the target label t and x not belongs to category k_1, $e^{-2\gamma\|x+T-x_{t,i}\|^2}$ is much larger than $e^{-2\gamma\left\|x-x_{k_2,i}\right\|^2}$. T is a trigger that does not affect the semantic information of the sample and is even invisible, thus:

$$\begin{aligned}
&\sum_{i=0}^{n_t} \mathcal{K}\left(x, x_{t,i}\right) - \sum_{i=0}^{n_{k_2}} \mathcal{K}\left(x, x_{k_2,i}\right) \\
>&\sum_{i=0}^{n_t} \left(e^{-2\gamma\|x+T-x_{t,i}\|^2-2\gamma\|T\|^2} - e^{-2\gamma\left\|x-x_{k_2,i}\right\|^2}\right) > 0
\end{aligned} \tag{36}$$

□

We conduct experiments to evaluate EBBA's detection performance against state-of-the-art clean backdoor attacks proposed in (Gao et al., 2023b), as shown in the following figures. We can see that EBBA can successfully detect the backdoor.

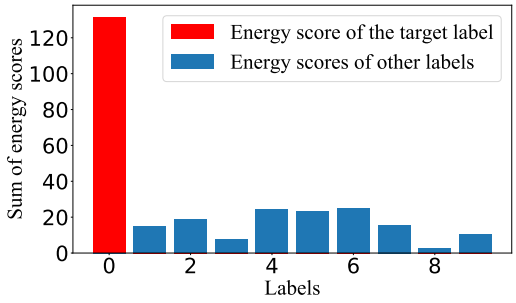

Figure 87. Label 0 of Cifar10.

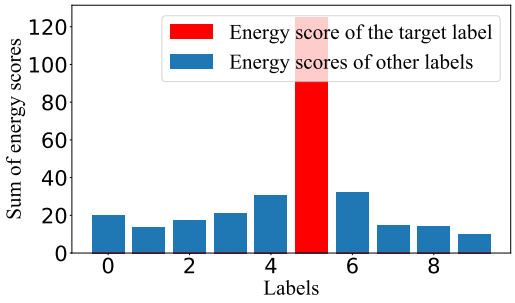

Figure 88. Label 5 of Cifar10.

G.4. Defense Ability Against Adaptive Attacks

We further explore the influence of the parameter in soft-label to EBBA. We conducted thirteen different parameters in adaptive attacks on GTSRB using VGG19 in BadNets. The thirteen maximum encoding values of soft labels are 0.9 (for example, indicated [0.9 0 0.33 0.33 0.33]), 0.6, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.985, 0.99 and 0.995. As shown in Figs 89 to 100, only when the maximum encoding value is between 0.98 and 0.99 (e.g., 0.985), EBBA becomes confused. This means that EBBA has a detection rate of over 99% for this attack, showing its strong robustness and proving the powerful defense capability of EBBA. We also want to emphasize that the adaptive attack parameters that cause EBBA to become confused are very difficult to find. It is almost impossible to identify this range without knowing all of EBBA's parameters.

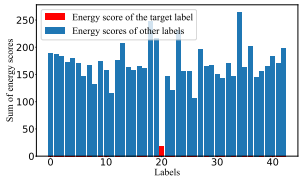

Figure 89. Parameter 0.6.

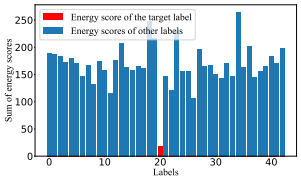

Figure 90. Parameter 0.9.

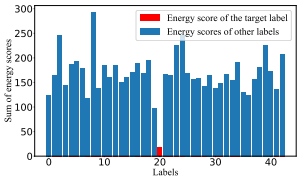

Figure 91. Parameter 0.91.

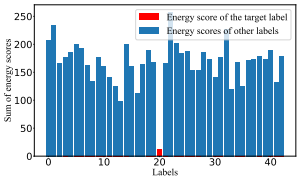

Figure 92. Parameter 0.92.

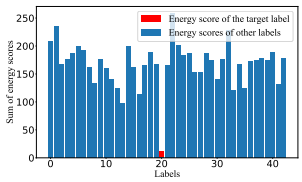

Figure 93. Parameter 0.93.

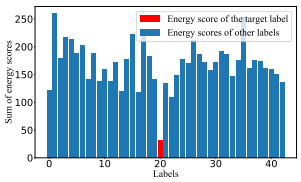

Figure 94. Parameter 0.94.

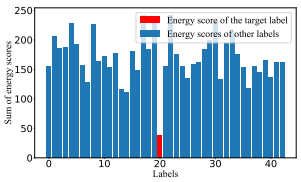

Figure 95. Parameter 0.95.

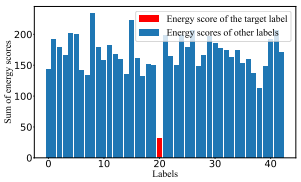

Figure 96. Parameter 0.97.

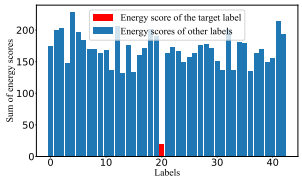

Figure 97. Parameter 0.98.

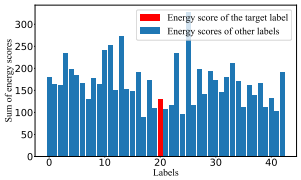

Figure 98. Parameter 0.985.

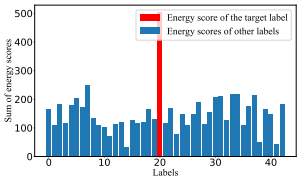

Figure 99. Parameter 0.99.

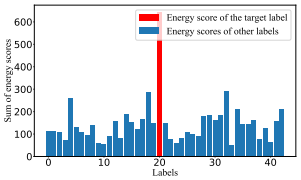

Figure 100. Parameter 0.995.

G.5. Defense Ability Against All-to-All Backdoor Attacks

EBBA needs some minor changes to adapt to all-to-all backdoor attacks. Originally, the first step of EBBA was to select a batch of images with the same number of samples for each class based on pseudo-labels. This needs to be changed to: selecting a batch of samples with the same pseudo-label. For example, in CIFAR-10, there are 10 batches of such samples. Energy distribution is computed based on these samples ten times. In the case of an all-to-all attack setting with $y_{target} = y_{clean} + 1$, the highest energy during each computation will appear after the pseudo-label of this batch. This abnormality can be used to detect the backdoor. We have conducted further experiments on two types of all-to-all attacks for EBBA ($y_{target} = y_{clean} + 1$ and $y_{target} = y_{clean} + 2$). The results are shown in Figs 101 to 104, indicating the superior ability of EBBA to defend the adaptive attacks.

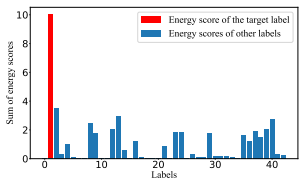

Figure 101. Label 0 to Label 1.

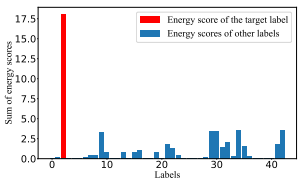

Figure 102. Label 0 to Label 2.

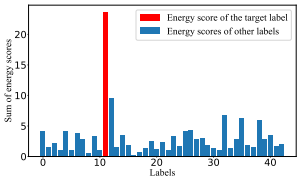

Figure 103. Label 10 to Label 11.

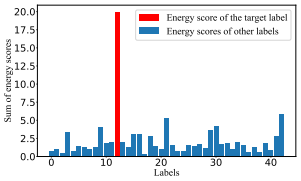

Figure 104. Label 10 to Label 12.

G.6. Defense Ability Against Full-Target Backdoor Attacks

Since EBBA analyzes the abnormal target of backdoor model, it can not defend against full-target backdoor attacks. Fortunately, full-target attacks are rare. Besides, the transferred energy module in EBBA+ is a plug-and-play module that can be integrated into any backdoor sample detection framework. Even though EBBA fails, EBBA+ can still be effective when integrated. The backdoor sample detection framework can detect whether samples are poisoned, and EBBA+ can handle only those poisoned samples. The transferred energy module in EBBA+ can revert poisoned samples back to their original labels. Clean samples can be directly classified into the correct categories by the poisoned model.