Fusing Pruned and Backdoored Models: Optimal Transport-based Data-free Backdoor Mitigation

Weilin Lin[1], **Li Liu**[1*], **Jianze Li**[2,3], **Hui Xiong**[1]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Shenzhen Research Institute of Big Data
[3]The Chinese University of Hong Kong, Shenzhen

Abstract

Backdoor attacks present a serious security threat to deep neuron networks (DNNs). Although numerous effective defense techniques have been proposed in recent years, they inevitably rely on the availability of either clean or poisoned data. In contrast, **data-free** defense techniques have evolved slowly and still lag significantly in performance. To address this issue, different from the traditional approach of pruning followed by fine-tuning, we propose a novel **data-free** defense method named ***Optimal Transport-based Backdoor Repairing*** (***OTBR***) in this work. This method, based on our findings on *neuron weight changes (NWCs)* of random unlearning, uses *optimal transport (OT)*-based model fusion to combine the advantages of both pruned and backdoored models. Specifically, we first demonstrate our findings that the NWCs of random unlearning are positively correlated with those of poison unlearning. Based on this observation, we propose a *random-unlearning NWC pruning* technique to eliminate the backdoor effect and obtain a backdoor-free pruned model. Then, motivated by the OT-based model fusion, we propose the *pruned-to-backdoored OT-based fusion* technique, which fuses pruned and backdoored models to combine the advantages of both, resulting in a model that demonstrates high clean accuracy and a low attack success rate. To our knowledge, this is the first work to introduce OT and model fusion techniques to the backdoor defense. Extensive experiments show that our method successfully defends against all seven backdoor attacks across three benchmark datasets, outperforming both state-of-the-art (SOTA) data-free and data-dependent methods.

Code — https://github.com/linweiii/OTBR
Extended version — https://arxiv.org/pdf/2408.15861

Introduction

Over the past decade, deep neural networks (DNNs) have become a crucial technology in various applications, including image recognition (Parmar and Mehta 2014; He et al. 2016a), speech processing (Gaikwad, Gawali, and Yannawar 2010; Maas et al. 2017), and natural language processing (Chowdhary and Chowdhary 2020), *etc*. However, as the deployment of DNNs in sensitive and critical domains becomes more widespread, concerns regarding their security cannot be ignored. Among the numerous threats to DNNs, *backdoor attacks* (Gu et al. 2019; Li et al. 2021a; Wu et al. 2023a) are particularly concerning. In these attacks, the attackers manipulate a small portion of the training data to implant a stealthy backdoor into a DNN, resulting in a *backdoored model*. During inference, the backdoored model behaves anomalously when the input contains a pre-defined trigger pattern; otherwise, it performs normally. This phenomenon is termed the *backdoor effect*. Such attacks may pose hidden security issues to real-world applications, such as unauthorized access to a system when a company develops its software using a third-party pre-trained model.

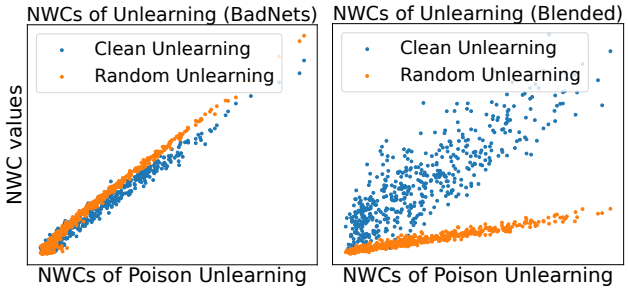

Figure 1: Illustration of unlearning NWCs on BadNets (Gu et al. 2019) and Blended (Chen et al. 2017) attacks. The NWCs of both clean and random unlearning show a positive correlation with poison unlearning. The last convolutional layer is chosen for this illustration.

In recent years, as backdoor attack methods have evolved, *backdoor defense* (Wu et al. 2023b) techniques have also seen significant growth. Various important techniques have been developed for backdoor defense, including pruning (Wu and Wang 2021; Zheng et al. 2022b), unlearning (Zeng et al. 2021a), and fine-tuning (Zhu et al. 2023), *etc*. However, most of these techniques rely on the availability of clean or poisoned data, which restricts their applicability to the aforementioned scenarios. Recent insights reveal a promising direction (Lin et al. 2024): using ***neuron weight changes (NWCs)*** of *clean unlearning*[1] to catego-

*Corresponds to Li Liu (avrillliu@hkust-gz.edu.cn)

[1]Unlearning the backdoored model on clean data.

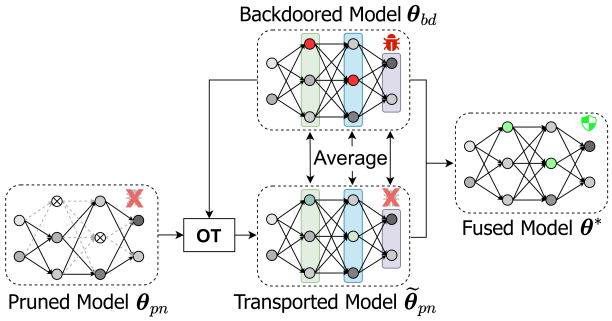

Figure 2: OT-based model fusion for backdoor defense. The pruned model is aligned with the backdoored model layer-by-layer using OT. Then the models are fused through a weighted averaging operation.

rize the neurons into backdoor-related ones and clean ones[2], based on an observation that the NWCs of unlearning clean and poisoned data are positively correlated. In this work, our extended findings reveal that using random noise for unlearning, termed as *random unlearning*, brings a new similar insight: **the NWCs of random unlearning exhibit a positive correlation with those of poison unlearning** (as shown in Figure 1). This motivates us to adopt NWCs for **data-free** backdoor mitigation using only the generated random noise. Normally, after identifying backdoor-related neurons, pruning followed by fine-tuning is employed to eliminate the backdoor effect and restore the lost performance (Liu, Dolan-Gavitt, and Garg 2018). However, this is infeasible in data-free scenarios since the subsequent fine-tuning requires clean data. If we only perform pruning using NWCs and simply skip the fine-tuning, the *clean accuracy* (ACC) is prone to decrease by more than 10% (Lin et al. 2024). Therefore, after pruning, it is necessary to develop a new data-free technique for performance recovery.

Recently, *model fusion* (Li et al. 2023a) has received increasing attention. It combines the weights of multiple models to integrate their capabilities into a single network. As one of the most representative works, OTFusion (Singh and Jaggi 2020) employs *optimal transport (OT)* to align model weights layer-by-layer before fusing two models through averaging. Following it, Intra-Fusion (Theus et al. 2024) employs OT to integrate the functionality of pruned neurons with the remaining ones, aiming to maintain great performance after pruning. It can be seen that the aforementioned methods both demonstrate OT's inherent ability to preserve critical information during the fusion process.

Motivated by the above advancements in model fusion, in this work, we explore its potential to combine the high ACC of the backdoored model with the low *attack success rate* (ASR) of the pruned model in a data-free manner. Building on the foundation of NWC pruning and OT-based model fusion, we propose a novel data-free defense strategy called ***O**ptimal **T**ransport-based **B**ackdoor **R**epairing* (OTBR), which fuses pruned and backdoored models. OTBR consists of two stages: *random-unlearning NWC pruning* and *pruned-to-backdoored OT-based fusion*. In the first stage, we calculate the NWCs based on random unlearning of the backdoored model, and then prune the top-ranking γ neurons to eliminate the backdoor effect. In the second stage, we align the weights of the pruned model with those of the backdoored model layer-by-layer using OT, and then fuse them into a single model. This process effectively dilutes the backdoor effect while preserving the clean performance. The fusion process is shown in Figure 2.

Our main contributions can be summarized as follows:

- We provide a new data-free pruning insight by revealing the positive correlation between NWCs when unlearning random noise and poisoned data.
- We propose a novel data-free defense strategy that combines the high ACC of the backdoored model with the low ASR of the pruned model, using the OT-based model fusion. To our knowledge, this is the first work to apply OT and model fusion techniques to backdoor defense.
- Experiments across various attacks, datasets, and experimental setups validate the effectiveness of our proposed OTBR method. Specifically, OTBR significantly outperforms both state-of-the-art (SOTA) data-free methods and SOTA data-dependent ones, consistently achieving successful defense performance against all tested attacks.

Related Work

Backdoor Attack

In the literature, various backdoor attacks on DNNs have been proposed, which can be generally categorized into two types: data-poisoning attacks and training-controllable attacks. For **data-poisoning attacks**, adversaries have access to the training dataset. BadNets (Gu et al. 2019), as one of the earliest examples, was proposed to implant a trigger pattern into the bottom-right corner of a small subset of the training images and reassign the labels to a specific target one. To enhance the stealthiness of the trigger, Blended (Chen et al. 2017) was proposed to blend the trigger onto the selected data with adjustable opacity. Recently, more sophisticated strategies have been proposed to enhance the trigger, including but not limited to SIG (Barni, Kallas, and Tondi 2019), label-consistent attacks (Shafahi et al. 2018; Zhao et al. 2020), and SSBA (Li et al. 2021a). Meanwhile, the second type, **training-controllable attacks**, is also rapidly evolving. In these attacks, adversaries have access to the training process, enabling more advanced attack strategies. Representative examples of this category include WaNet (Nguyen and Tran 2021) and Input-aware (Nguyen and Tran 2020), which incorporate an injection function into the training process to generate unique triggers for each input data. These innovative tactics make it more challenging to detect the triggers and conduct an effective defense.

Backdoor Defense

In general, backdoor defense methods can be categorized into three types: pre-training, in-training, and post-training defenses. Among them, **Post-training Defense** has received the most attention, where the defenders aim to mitigate

[2]The larger NWC of a neuron, the more backdoor-related it is.

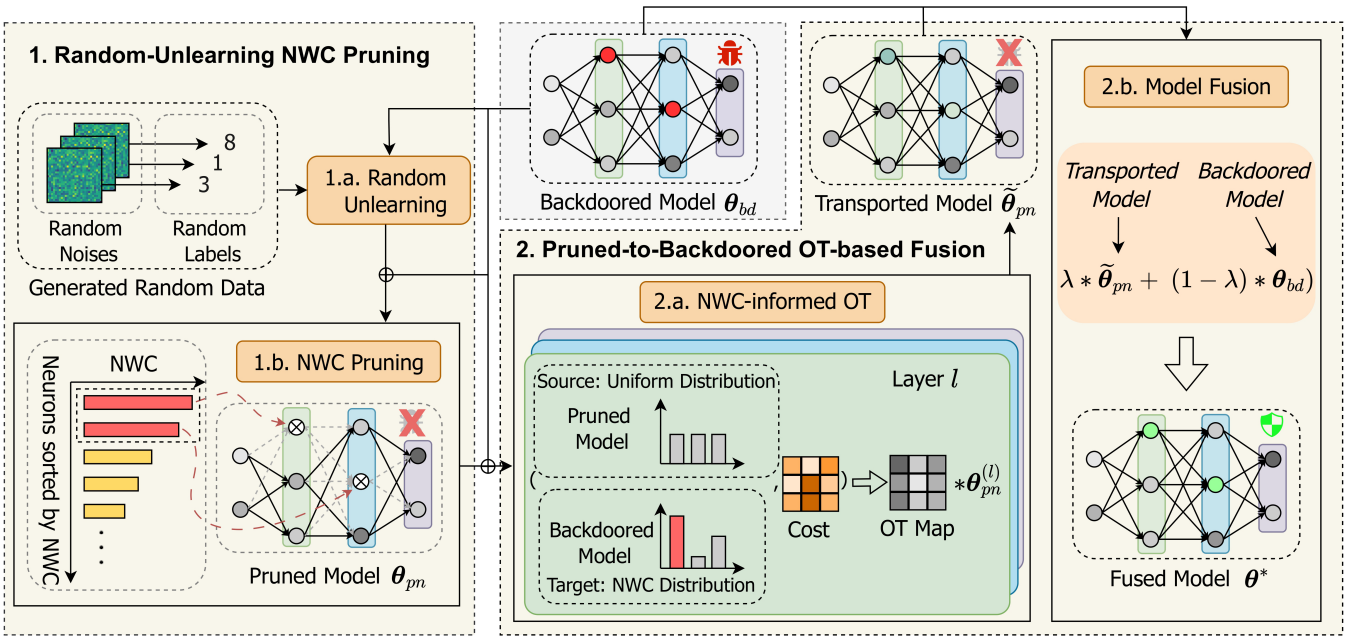

Figure 3: Overview of the proposed OTBR framework.

the backdoor effect of a well-trained backdoored model. FP (Liu, Dolan-Gavitt, and Garg 2018), as one of the seminal defense methods, prunes the less-activated neurons and then fine-tunes the model, based on the observation that poisoned and clean data activate different neurons; ANP (Wu and Wang 2021) detects and prunes backdoor-related neurons by applying adversarial perturbations to neuron weights; Building on this, RNP (Li et al. 2023b) refines the perturbation technique using clean unlearning, and performs pruning based on a learned mask. In addition to these pruning-based techniques, some other important defense techniques exist. NC (Wang et al. 2019) proposes recovering the trigger to improve backdoor removal; NAD (Li et al. 2021c) pioneers the use of model distillation to train a benign student model; i-BAU (Zeng et al. 2021a) uses adversarial attacks to identify potential triggers and then performs poison unlearning to mitigate the backdoor effect.

Different from the above defenses, which are all data-dependent, CLP (Zheng et al. 2022a) is the first data-free defense method, which identifies and prunes potential backdoored neurons based on channel Lipschitzness; ABD (Hong et al. 2023) designs a plug-in defensive technique specialized for data-free knowledge distillation; DHBE (Yan et al. 2023) proposes a competing strategy between distillation and backdoor regularization to distill a clean student network without data.

Although several data-dependent techniques have already been proposed in the literature, the scarcity of data-free defense techniques still limits the applicability of backdoor defenses in real-world scenarios. In this paper, we will focus on addressing this issue, and develop a novel effective data-free defense method by using random-unlearning NWCs and the OT-based model fusion technique.

Preliminary

Threat Model

In this work, we address threats from both data-poisoning and training-controllable attacks. The attackers aim to poison a small portion of the training data so that the trained model predicts a target class when presented with data containing a pre-defined *trigger*, while otherwise performing normally. The weights of a L-layer backdoored model are denoted as $\boldsymbol{\theta}_{bd} = \{\boldsymbol{\theta}_{bd}^{(l)}\}_{1 \le l \le L}$, where $\boldsymbol{\theta}_{bd}^{(l)}$ represents the weights for the l^{th} layer, consisting of $m^{(l)}$ neurons.

Defense Setting

We focus on the post-training scenario, aiming to mitigate the backdoor effect of a well-trained backdoored model while minimizing the negative impact on ACC. Different from most previous works (Liu, Dolan-Gavitt, and Garg 2018; Wu and Wang 2021; Zeng et al. 2021a), which assumes access to 5% of clean data for defense, we adopt a more stringent approach that relies only on the backdoored model, without access to any clean data (Zheng et al. 2022a).

Method

Overview of Our Method

The complete data-free defense process of our proposed OTBR strategy is illustrated in Figure 3, which consists of two stages as follows:

- In **Stage 1**, referred to as *random-unlearning NWC pruning*, we aim to obtain a backdoor-free model. Specifically, we first conduct random unlearning on the backdoored model for I iterative steps. During each step, a

mini-batch of random noise with random labels is generated and used for unlearning. Then, we calculate the NWC for each neuron based on their weights from both the backdoored and unlearned models. Finally, we prune the top-ranking γ of neurons, based on their NWCs, from the backdoored model to eliminate its backdoor effect.

- In **Stage 2**, referred to as *pruned-to-backdoored OT-based fusion*, when obtaining a sub-optimal pruned model, we aim to combine its low ASR with the high ACC of the original backdoored model by repairing the backdoor-related neurons using OT-based model fusion. Specifically, we propose *NWC-informed OT* to align the weights of the pruned model with those of the backdoored model layer-by-layer, taking into account the backdoor importance as determined by NWCs. For each layer, starting with the earliest pruned one, we initialize the probability mass on neuron weights using a uniform distribution for the pruned model and an NWC distribution for the original backdoored model. This strategy discriminatively transfers clean functionality to the backdoor-related neurons. The cost matrix $\boldsymbol{C}^{(l)}$ is calculated based on the Euclidean distance between neuron weights from the two models. Using this cost matrix, we then derive the optimal transport map $\mathbf{T}^{(l)}$ and employ it to transport the weights of the pruned model. After aligning all layers, we perform a simple weight averaging to fuse the transported and backdoored models, resulting in an effective defense.

Next, we will present more detailed formulations and provide further insights.

Stage 1: Random-Unlearning NWC Pruning

Random Unlearning. Unlearning is a reverse training process designed to maximize the loss value on a given dataset (Li et al. 2023b). In this work, we define *random unlearning* as the process of unlearning a DNN model f using a generated random dataset $\mathcal{D}_r$. More precisely, random unlearning on the backdoored model $\boldsymbol{\theta}_{bd}$ is formulated as:

$$\max_{\boldsymbol{\theta}_{bd}} \mathbb{E}_{(\boldsymbol{x}_r, y_r)\in\mathcal{D}_r} \left[\mathcal{L}(f(\boldsymbol{x}_r; \boldsymbol{\theta}_{bd}), y_r)\right], \tag{1}$$

where the loss function $\mathcal{L}$ is chosen to be a cross-entropy loss, and the generated random dataset $\mathcal{D}_r$ contains $I \times B$ pairs of random noises $\boldsymbol{x}_r \in [0,1]^{A\times H\times W}$ and random labels $y_r \in \{0, 1, \ldots, G\}$. Here, I is the number of iterative steps; B is the batch size; A, H and W represent the generated noise size; and G is the largest class label.

NWC Pruning. We follow the NWC definition from (Lin et al. 2024) to quantify the weight changes for each neuron during unlearning. Specifically, for the j-th neuron in the l-th layer, the NWC is defined as:

$$\mathrm{NWC}^{(l)j} \stackrel{\mathsf{def}}{=} \|\boldsymbol{\theta}_{ul}^{(l)j} - \boldsymbol{\theta}_{bd}^{(l)j}\|_1, \tag{2}$$

where $\boldsymbol{\theta}_{ul}$ denotes the unlearned backdoored model, $j \in \{1, \ldots, m^{(l)}\}$ and $l \in \{1, \ldots, L\}$. To eliminate the backdoor effect, we sort all calculated NWCs in descending order and prune the top-ranking γ of neurons from the original backdoored model. The pruned model is denoted as $\boldsymbol{\theta}_{pn}$.

Stage 2: Pruned-to-Backdoored OT-based Fusion

Optimal Transport. OT is a mathematical framework to find the most economical way to transport mass from one distribution to another. Suppose we have two discrete probability distributions in the space $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\mathcal{Y} = \{y_j\}_{j=1}^m$, *i.e.*, the source distribution $\mu := \sum_{i=1}^n \alpha_i \cdot \delta(x_i)$ and the target distribution $\nu := \sum_{j=1}^m \beta_j \cdot \delta(y_j)$, where $\sum_{i=1}^n \alpha_i = \sum_{j=1}^m \beta_j = 1$ and $\delta(\cdot)$ is the Dirac delta function. The OT problem can be formulated as a linear programming problem as follows:

$$\begin{aligned} \mathrm{OT}(\mu, \nu; \boldsymbol{C}) &\stackrel{\mathsf{def}}{=} \min\langle \mathbf{T}, \boldsymbol{C}\rangle, \\ \text{s.t., } \mathbf{T}\mathbf{1}_m &= \boldsymbol{\alpha},\ \mathbf{T}^\top \mathbf{1}_n = \boldsymbol{\beta} \end{aligned}, \tag{3}$$

where $\mathbf{T} \in \mathbb{R}_+^{n\times m}$ is the transport map that determines the optimal transport amount of mass from $\mathcal{X}$ to $\mathcal{Y}$, and $\boldsymbol{C}$ is the cost matrix quantifying the cost of moving each unit of mass.

NWC-informed OT. In our approach, we align the NWC-pruned model, which contains only clean functionality, with the original backdoored model to achieve more effective fusion. The goal is to dilute the backdoor effect with the least influence on clean performance. To achieve this, we focus more on the backdoor-related neurons during weight transport by employing NWC-informed initialization for the target distribution ν (backdoored model), while using a uniform distribution for the source distribution μ (pruned model). For the l-th layer, we denote the probability mass as:

$$\boldsymbol{\alpha}^{(l)} \stackrel{\mathsf{def}}{=} \left\{\frac{1}{n^{(l)}}\right\}_{i=1}^{n^{(l)}},\ \boldsymbol{\beta}^{(l)} \stackrel{\mathsf{def}}{=} \left\{\frac{\mathrm{NWC}^{(l)j}}{\sum_{j=1}^{m^{(l)}} \mathrm{NWC}^{(l)j}}\right\}_{j=1}^{m^{(l)}}, \tag{4}$$

where $n^{(l)}$ and $m^{(l)}$ denote the neuron numbers of pruned and backdoored models, respectively. Then, based on the distributions $\mu^{(l)}$ and $\nu^{(l)}$, and the cost matrix $\boldsymbol{C}^{(l)}$, we can derive the optimal transport map $\mathbf{T}^{(l)}$ by equation (3).

Model Fusion. Inspired by the OTFusion (Singh and Jaggi 2020), we align and fuse the pruned and backdoored models layer-by-layer, using OT in equation (3) and our defined distributions in equation (4). The details of the entire fusion process are shown in Algorithm 1. Note that we start the fusion process from the first pruned layer p, rather than the second layer. For the l-th layer, $n^{(l)}$ and $m^{(l)}$ represent the neuron number of $\boldsymbol{\theta}_{pn}^{(l)}$ and $\boldsymbol{\theta}_{bd}^{(l)}$, respectively.

In Algorithm 1, for each layer l, we first align the incoming edge weights using the OT map $\mathbf{T}^{(l-1)}$ and probability mass $\boldsymbol{\beta}^{(l-1)}$ from the previous layer:

$$\widehat{\boldsymbol{\theta}}_{pn}^{(l)} \leftarrow \boldsymbol{\theta}_{pn}^{(l)} \mathbf{T}^{(l-1)} \mathrm{diag}(1/\boldsymbol{\beta}^{(l-1)}).$$

Then, we get the distributions $\mu^{(l)}$ and $\nu^{(l)}$ of the current layer and compute the cost matrix $\boldsymbol{C}^{(l)}$ using Euclidean distance between neuron weights: $\boldsymbol{C}_{ij}^{(l)} \stackrel{\mathsf{def}}{=} \|\boldsymbol{\theta}_{pn}^{(l)i} - \boldsymbol{\theta}_{bd}^{(l)j}\|^2$. Finally, as in equation (3), using $\mu^{(l)}$, $\nu^{(l)}$ and $\boldsymbol{C}^{(l)}$, the OT

Algorithm 1: Pruned-to-Backdoored OT-based Fusion

Input: Pruned model $\boldsymbol{\theta}_{pn}$, backdoored model $\boldsymbol{\theta}_{bd}$, random-unlearning NWC values for each neuron, balance coefficient λ, the first pruned layer p.
Output: Clean model $\boldsymbol{\theta}^*$.

$\boldsymbol{\alpha}^{(p-1)} \leftarrow \left\{1/n^{(p-1)}\right\}_{i=1}^{n^{(p-1)}}$
$\boldsymbol{\beta}^{(p-1)} \leftarrow \left\{1/m^{(p-1)}\right\}_{j=1}^{m^{(p-1)}}$
$\mathbf{T}^{(p-1)} \leftarrow \text{diag}(\boldsymbol{\beta}^{(p-1)})\mathbf{I}_{m^{(p-1)}\times m^{(p-1)}}$
for $l = p$ **to** L **do**
  $\widehat{\boldsymbol{\theta}}_{pn}^{(l)} \leftarrow \boldsymbol{\theta}_{pn}^{(l)}\mathbf{T}^{(l-1)}\text{diag}(1/\boldsymbol{\beta}^{(l-1)})$
  $\boldsymbol{\alpha}^{(l)} \leftarrow \left\{1/n^{(l)}\right\}_{i=1}^{n^{(l)}}$
  $\boldsymbol{\beta}^{(l)} \leftarrow \left\{\text{NWC}^{(l)j} / \sum_{j=1}^{m^{(l)}} \text{NWC}^{(l)j}\right\}_{j=1}^{m^{(l)}}$
  $\mu^{(l)}, \nu^{(l)} \leftarrow \text{GetDistribution}(\boldsymbol{\alpha}^{(l)}, \boldsymbol{\beta}^{(l)})$
  $\boldsymbol{C}^{(l)} \leftarrow \text{ComputeCost}(\boldsymbol{\theta}_{pn}^{(l)}, \boldsymbol{\theta}_{bd}^{(l)})$
  $\mathbf{T}^{(l)} \leftarrow \text{OT}(\mu^{(l)}, \nu^{(l)}, \boldsymbol{C}^{(l)})$
  $\widetilde{\boldsymbol{\theta}}_{pn}^{(l)} \leftarrow \text{diag}(1/\boldsymbol{\beta}^{(l)})\mathbf{T}^{(l)\top}\widehat{\boldsymbol{\theta}}_{pn}^{(l)}$
  $\boldsymbol{\theta}^{*(l)} \leftarrow \lambda\widetilde{\boldsymbol{\theta}}_{pn}^{(l)} + (1-\lambda)\boldsymbol{\theta}_{bd}^{(l)}$
end for
Obtain clean model $\boldsymbol{\theta}^*$

map $\mathbf{T}^{(l)}$ of the current layer can be derived, and the transported pruned model $\widetilde{\boldsymbol{\theta}}_{pn}^{(l)}$ can be obtained as:

$$\widetilde{\boldsymbol{\theta}}_{pn}^{(l)} \leftarrow \text{diag}\left(1/\boldsymbol{\beta}^{(l)}\right)\mathbf{T}^{(l)\top}\widehat{\boldsymbol{\theta}}_{pn}^{(l)},$$

which has been aligned with the backdoored model.

After alignment, the transported model is then fused with the backdoored model to obtain the final defense model, which can be formulated as:

$$\boldsymbol{\theta}^* \leftarrow \lambda\widetilde{\boldsymbol{\theta}}_{pn} + (1-\lambda)\boldsymbol{\theta}_{bd}, \tag{5}$$

where $\boldsymbol{\theta}^*$ represents the fused clean model and λ is the balance coefficient.

Why Can OT-based Fusion Mitigate Backdoor Effect? We now offer a possible explanation for the effectiveness of OT-based model fusion in mitigating backdoor effects. Based on previous work (Lin et al. 2024), the NWC-pruned model can be made backdoor-free, *i.e.*, the ASR dropping to zero, by selecting a suitable pruning threshold. Therefore, by aligning the pruned model with the original backdoored model using NWC-informed OT, we can transport the clean functionality of the remaining neurons to the nearest backdoored positions, as determined by NWCs and Euclidean distance. Then, further fusion based on the transported model can be viewed as a dilution operation to weaken the backdoored effect of the original backdoored model while preserving its clean functionality, thanks to the inherent ability of OT (Singh and Jaggi 2020; Theus et al. 2024). This is consistent with the previous insights that the backdoor task is easier and encoded in much fewer neurons than the clean task (Li et al. 2021b; Cai et al. 2022).

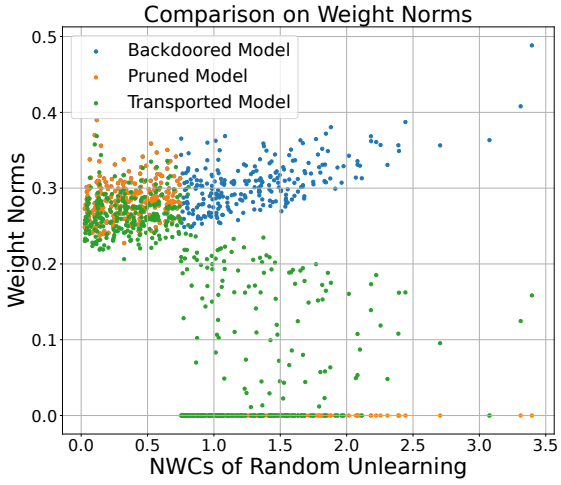

Figure 4: Illustration of neuron-level weight norms for the backdoored, pruned, and transported models during OT-based fusion.

A practical example of the BadNets-attacked PreAct-ResNet18 (He et al. 2016b) is illustrated in Figure 4. From the perspective of *weight norm*, the larger the difference in a neuron's weight between the pruned and transported models, the more it is transported by the OT. We observe a consistent, slight decrease in the unpruned neurons that are intensively transported to some specific pruned neurons, resulting in their rapid recovery. This outcome reflects the effect of transporting from a uniform source distribution to a NWC-informed target distribution. By fusing the transported (green) and backdoored (blue) models, we can discriminately modify the neuron functionality, *e.g.*, recovering more in low-NWC neurons, to effectively mitigate the backdoor effect while preserving high performance.

Experiment

Experimental Setup

For a fair comparison, all experiments, including the code implementation of our proposed method, are conducted using the default settings in BackdoorBench (Wu et al. 2022).

Datasets. Similar to previous works (Zhu et al. 2023; Wei et al. 2023), our experiments are conducted on three benchmark datasets, including CIFAR-10 (Krizhevsky, Hinton et al. 2009), Tiny ImageNet (Le and Yang 2015), and CIFAR-100 (Krizhevsky, Hinton et al. 2009).

Attack Setup. We evaluate the effectiveness of all defense methods using seven SOTA backdoor attacks: BadNets (Gu et al. 2019), Blended (Chen et al. 2017), Input-aware (Nguyen and Tran 2020), LF (Zeng et al. 2021b), SSBA (Li et al. 2021a), Trojan (Liu et al. 2018) and WaNet (Nguyen and Tran 2021). All attacks are conducted using the default settings in BackdoorBench (Wu et al. 2022). For example, we set the target label to 0, the poisoning ratio to 10%, and the tested model to PreAct-ResNet18 (He et al. 2016b).

Datasets	Attacks	No Defense		Data-Dependent										Data-Free			
				FP		NAD		ANP		i-BAU		RNP		CLP		**OTBR**	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR-10	BadNets	91.32	95.03	91.31	57.13	89.87	2.14	90.94	5.91	89.15	1.21	89.81	24.97	90.06	77.50	90.11	1.08
	Blended	93.47	99.92	93.17	99.26	92.17	97.69	93.00	84.90	87.00	50.53	88.76	79.74	91.32	99.74	92.01	1.64
	Input-aware	90.67	98.26	91.74	0.04	93.18	1.68	91.04	1.32	89.17	27.08	90.52	1.84	90.30	2.17	86.52	0.37
	LF	93.19	99.28	92.90	98.97	92.37	47.83	92.83	54.99	84.36	44.96	88.43	7.02	92.84	99.18	87.68	9.69
	SSBA	92.88	97.86	92.54	83.50	91.91	77.40	92.67	60.16	87.67	3.97	88.60	17.89	91.38	68.13	85.27	9.54
	Trojan	93.42	100.00	92.46	71.17	91.88	3.73	92.97	46.27	90.37	2.91	90.89	3.59	92.98	100.00	90.62	7.50
	WaNet	91.25	89.73	91.46	1.09	93.17	22.98	91.32	2.22	89.49	5.21	90.43	0.96	81.91	78.42	88.12	10.93
	Average	92.31	97.15	**92.23**	58.74	92.08	36.21	92.11	36.54	88.17	19.41	89.63	19.43	90.11	75.02	88.62	**5.82**
Tiny ImageNet	BadNets	56.23	100.00	51.73	99.99	46.37	0.27	50.55	7.74	51.48	97.36	21.91	0.00	55.94	100.00	54.13	0.00
	Input-aware	57.45	98.85	55.28	62.92	47.91	1.86	53.17	0.17	52.48	72.98	15.57	0.00	57.75	99.58	51.40	0.02
	SSBA	55.22	97.71	50.47	88.87	45.32	57.32	52.83	91.44	49.86	81.90	37.64	0.00	55.17	97.65	54.15	3.89
	Trojan	55.89	99.98	50.22	8.82	48.48	0.83	50.37	1.40	52.65	98.49	46.27	0.00	55.86	8.39	53.85	0.14
	WaNet	56.78	99.49	53.84	3.94	46.98	0.43	53.87	0.75	53.71	75.23	20.50	0.00	56.21	98.50	55.64	0.03
	Average	56.31	99.21	52.31	52.91	47.01	12.14	52.16	20.30	52.04	85.19	28.38	**0.00**	**56.19**	80.82	53.83	0.82
CIFAR-100	BadNets	67.22	87.43	64.55	0.42	66.37	0.06	63.65	0.00	60.37	0.04	55.68	0.00	65.40	81.95	66.81	0.00
	Input-aware	65.24	98.61	67.82	2.34	69.25	31.11	58.99	0.00	65.21	85.14	55.66	0.01	65.22	99.81	59.64	4.21
	SSBA	69.06	97.22	61.60	14.02	67.38	89.51	64.35	39.60	63.09	28.91	68.44	92.80	65.39	97.52	66.89	1.28
	WaNet	64.04	97.72	68.07	10.29	68.46	0.55	60.05	0.05	65.31	43.96	49.48	0.00	25.90	83.49	62.91	8.18
	Average	66.39	95.25	65.51	6.77	**67.87**	30.31	61.76	9.91	63.50	39.51	57.32	23.20	55.48	90.69	64.06	**3.42**
Average ACC Drop (smaller is better)				↓**1.51**	-	↓2.64	-	↓2.55	-	↓3.87	-	↓12.17	-	↓3.73	-	↓2.97	-
Average ASR Drop (larger is better)				-	↓53.39	-	↓70.11	-	↓72.51	-	↓52.33	-	↓83.02	-	↓16.57	-	↓**93.66**
Successful Defense Count				8 / 16		9 / 16		10 / 16		5 / 16		7 / 16		2 / 16		**16** / 16	

Table 1: Performance comparison with the SOTA defenses on CIFAR-10, Tiny ImageNet, and CIFAR-100 (%).

Defense Setup. We compare our proposed OTBR method with six SOTA defense methods: Fine-pruning (FP) (Liu, Dolan-Gavitt, and Garg 2018), NAD (Li et al. 2021c), ANP (Wu and Wang 2021), i-BAU (Zeng et al. 2021a), RNP (Li et al. 2023b), and CLP (Zheng et al. 2022a). Note that only CLP can be conducted in a data-free manner similar to OTBR, while the other five defenses are all data-dependent. Therefore, we follow the common setting in the post-training scenario that 5% clean data is provided for those methods.

Evaluation Metrics. We use two common metrics to evaluate performance: ACC and ASR. They measure the proportion of correct predictions on clean data (the higher, the better) and the rate of incorrect predictions for the target label on poisoned data (the lower, the better), respectively. A defense is usually considered successful against an attack if the ASR is reduced to below 20% (Qi et al. 2023; Xie et al. 2024). In this paper, to consider both ACC and ASR, we consider a defense to be **successful** (marked with green in all tables) only if it achieves both of the following criteria: ACC decreases by less than 10% and ASR falls below 20%. Otherwise, it is considered unsuccessful. The best average results are **boldfaced** in all tables within this section.

Compared with Previous Works

The main defense performance, compared with the six baseline methods, is shown in Table 1. We observe that our OTBR successfully defends against all 16 attacks across three benchmark datasets, achieving the largest drop in average ASR (93.66%) with an acceptable average ACC reduction (2.97%). Moreover, OTBR outperforms both data-free and data-dependent defenses, achieving the lowest average ASR on CIFAR-10 and CIFAR-100, and the second-best ASR on Tiny ImageNet. Notably, the best ASR on Tiny ImageNet, achieved by RNP, comes with a significant drop in ACC. For the performances of baseline methods, we observe that the data-dependent approaches have clear advantages over the data-free CLP, which succeeds in only 2 out of 16 defenses. ANP achieves the best results with 10 out of 16 successful defenses; however, it consistently fails against SSBA attacks, a shortcoming also observed in NAD. Despite its failures on CIFAR-10, FP performs well on CIFAR-100 and consistently achieves high ACCs across all three datasets, validating the effectiveness of fine-tuning. i-BAU fails completely on Tiny ImageNet, exposing its limitations when dealing with different data complexities. Although RNP achieves good performance in ASR, it fails with significant drops in ACC on Tiny ImageNet. In contrast, OTBR performs the best, consistently achieving superior results across various attacks and datasets.

Attacks	$\mathbf{V}_1$		$\mathbf{V}_2$		$\mathbf{V}_3$ (Ours)	
	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	84.98	1.56	91.3	81.13	90.11	1.08
Blended	69.08	4.71	92.06	16.76	92.01	1.64
LF	54.15	0.83	90.97	60.4	87.68	9.69
SSBA	40.58	0.02	89.84	40.57	85.27	9.54

Table 2: Comparison of different fusion schemes (%). $\mathbf{V}_1$: no fusion; $\mathbf{V}_2$: vanilla fusion; $\mathbf{V}_3$: OT-based fusion.

Ablation Studies

Effectivness of OT-based Model Fusion. To evaluate the effectiveness of OT-based model fusion, we keep Stage 1 unchanged and modify Stage 2 to generate three different versions for comparison. (1) $\mathbf{V}_1$: no fusion is conducted in Stage 2; instead, we evaluate the performance of the pruned model from Stage 1; (2) $\mathbf{V}_2$: implement vanilla fusion by directly fusing the pruned model with the backdoored model using equation (5); (3) $\mathbf{V}_3$ (Ours): the final version, where the full procedures of both Stage 1 and 2 are conducted. Table 2 shows the performances of these three versions on CIFAR-10 across four different attacks. The results validate the ef-

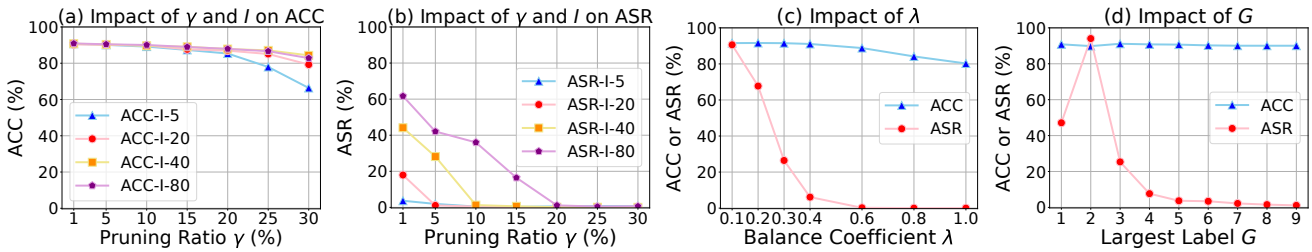

Figure 5: Impact of different factors on performance. (a) and (b) show the impact of γ and I on ACC and ASR, respectively, with "ACC-I-5" representing ACC when $I = 5$; (c) shows the impact of λ; and (d) shows the impact of G.

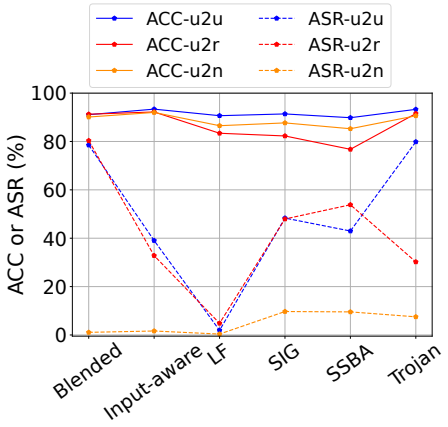

Figure 6: Comparison of different OT distributions. u2u: uniform to uniform transport; u2r: uniform to random transport; u2n(ours): uniform to NWC transport.

fectiveness of aligning neuron weights using OT, where the pruned model inherently achieves a low ASR (or even better) while the high ACC is kept. Although the vanilla fusion ($\mathbf{V}_2$) better preserves ACC, it fails in effectively mitigating the backdoor effect.

Effectiveness of NWC-informed OT. To verify the important role of NWC-informed OT in achieving optimal defense performance, we fix the source distribution as a uniform distribution and compare three different initialization schemes for the target distribution in OT. Specifically, we consider three types of distributions: *uniform distribution, random distribution*, and *NWC distribution*. These schemes are labeled as "u2u", "u2r", and "u2n"(Ours), respectively. The results are presented in Figure 6. We observe that only "u2n", *i.e.*, NWC-informed OT, consistently achieves strong performance across different attacks. It can be explained *w.r.t.* the weight transport as in Figure 4. In contrast, since uniform and random distributions are unrelated to the backdoor functionality, they fail to effectively guide neuron weight transport, resulting in poor fusion performance.

Parameter Analysis

Impact of Different Factors. We aim to investigate the impact of various factors on the performance of OTBR. These factors include the number of iterative steps I, the pruning ratio γ, the largest label G, and the balance coefficient λ. The experiments are conducted using default settings on CIFAR-10 and BadNets with a 10% poisoning ratio. The results are shown in **Figure 5**. **Firstly**, in subfigures (a) and (b), we present the results of varying the pruning ratio γ from 1% to 30% across four numbers of iterative steps I (5, 20, 40, and 80 steps). The two subfigures, showing ACC and ASR respectively, demonstrate that OTBR performs well across different settings. A larger pruning ratio tends to require more iterative steps for effective random unlearning. In our setting ($I = 20$), γ is insensitive in the range of 5% to 25%, resulting in successful defense. **Secondly**, in subfigure (c), we show the impact of the balance coefficient λ ranging from 0.1 to 1.0. A higher value of λ means a more important role the transported model plays in the fusion. We observe that there exists a trade-off between the high ACC from the backdoored model and the low ASR from the transported model, as we assumed before. It suggests setting the λ between 0.4 and 0.8 for a successful defense. **Lastly**, in subfigure (d), we evaluate the performances with different largest labels G for random data generation to test the impact of class number. Note that 9 is the largest label, in which case the model is trained. The results reveal that performance remains consistently good across G values from 4 to 9, while a smaller class number may fail. Overall, our OTBR proves to be a robust defense method across various hyperparameter settings.

Conclusion

In this work, we propose a novel data-free backdoor defense method, OTBR, using OT-based model fusion. Notably, we provide a new data-free pruning insight by revealing the positive correlation between NWCs when unlearning random noise and poisoned data. This insight enables us to effectively eliminate the backdoor effect using pruning guided by NWCs in a data-free manner. Then, we propose to combine the high ACC of the backdoored model with the low ASR of the pruned model using the OT-based model fusion. Furthermore, we provide possible explanations for the success of both NWC pruning and OT-based fusion. Extensive experiments across various attacks and datasets confirm the effectiveness of our OTBR method. A current limitation of this work is its reliance on NWC, which applies only to the post-training scenario. In future work, we plan to explore the potential of OT-based model fusion for more scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62471420 and 62101351), Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2023A03J0008), and Education Bureau of Guangzhou Municipality.

References

Barni, M.; Kallas, K.; and Tondi, B. 2019. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, 101–105. IEEE.

Cai, R.; Zhang, Z.; Chen, T.; Chen, X.; and Wang, Z. 2022. Randomized channel shuffling: Minimal-overhead backdoor attack detection without clean datasets. *Advances in Neural Information Processing Systems*, 35: 33876–33889.

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

Chowdhary, K.; and Chowdhary, K. 2020. Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Gaikwad, S. K.; Gawali, B. W.; and Yannawar, P. 2010. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3): 16–24.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 630–645. Springer.

Hong, J.; Zeng, Y.; Yu, S.; Lyu, L.; Jia, R.; and Zhou, J. 2023. Revisiting data-free knowledge distillation with poisoned teachers. In *International Conference on Machine Learning*, 13199–13212. PMLR.

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

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, W.; Peng, Y.; Zhang, M.; Ding, L.; Hu, H.; and Shen, L. 2023a. Deep model fusion: A survey. *arXiv preprint arXiv:2309.15698*.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021a. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16463–16472.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021b. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34: 14900–14912.

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

Li, Y.; Lyu, X.; Ma, X.; Koren, N.; Lyu, L.; Li, B.; and Jiang, Y.-G. 2023b. Reconstructive neuron pruning for backdoor defense. In *International Conference on Machine Learning*, 19837–19854. PMLR.

Lin, W.; Liu, L.; Wei, S.; Li, J.; and Xiong, H. 2024. Unveiling and Mitigating Backdoor Vulnerabilities based on Unlearning Weight Changes and Backdoor Activeness. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, 273–294. Springer.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.-C.; Zhai, J.; Wang, W.; and Zhang, X. 2018. Trojaning attack on neural networks. In *NDSS Symposium*.

Maas, A. L.; Qi, P.; Xie, Z.; Hannun, A. Y.; Lengerich, C. T.; Jurafsky, D.; and Ng, A. Y. 2017. Building DNN acoustic models for large vocabulary speech recognition. *Computer Speech & Language*, 41: 195–213.

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

Nguyen, T. A.; and Tran, A. 2020. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33: 3454–3464.

Parmar, D. N.; and Mehta, B. B. 2014. Face recognition methods & applications. *arXiv preprint arXiv:1403.0485*.

Qi, X.; Xie, T.; Wang, J. T.; Wu, T.; Mahloujifar, S.; and Mittal, P. 2023. Towards a proactive {ML} approach for detecting backdoor poison samples. In *32nd USENIX Security Symposium (USENIX Security 23)*, 1685–1702.

Shafahi, A.; Huang, W. R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; and Goldstein, T. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31.

Singh, S. P.; and Jaggi, M. 2020. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33: 22045–22055.

Theus, A.; Geimer, O.; Wicke, F.; Hofmann, T.; Anagnostidis, S.; and Singh, S. P. 2024. Towards Meta-Pruning via Optimal Transport. In *The Twelfth International Conference on Learning Representations*.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, 707–723. IEEE.

Wei, S.; Zhang, M.; Zha, H.; and Wu, B. 2023. Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples. *Advances in Neural Information Processing Systems*, 36: 25876–25909.

Wu, B.; Chen, H.; Zhang, M.; Zhu, Z.; Wei, S.; Yuan, D.; and Shen, C. 2022. Backdoorbench: A comprehensive benchmark of backdoor learning. In *NeurIPS Datasets and Benchmarks Track*.

Wu, B.; Liu, L.; Zhu, Z.; Liu, Q.; He, Z.; and Lyu, S. 2023a. Adversarial machine learning: A systematic survey of backdoor attack, weight attack and adversarial example. *arXiv preprint arXiv:2302.09457*, 1.

Wu, B.; Wei, S.; Zhu, M.; Zheng, M.; Zhu, Z.; Zhang, M.; Chen, H.; Yuan, D.; Liu, L.; and Liu, Q. 2023b. Defenses in adversarial machine learning: A survey. *arXiv preprint arXiv:2312.08890*.

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

Xie, T.; Qi, X.; He, P.; Li, Y.; Wang, J. T.; and Mittal, P. 2024. BaDExpert: Extracting Backdoor Functionality for Accurate Backdoor Input Detection. In *The Twelfth International Conference on Learning Representations*.

Yan, Z.; Li, S.; Zhao, R.; Tian, Y.; and Zhao, Y. 2023. DHBE: data-free holistic backdoor erasing in deep neural networks via restricted adversarial distillation. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, 731–745.

Zeng, Y.; Chen, S.; Park, W.; Mao, Z. M.; Jin, M.; and Jia, R. 2021a. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*.

Zeng, Y.; Park, W.; Mao, Z. M.; and Jia, R. 2021b. Rethinking the backdoor attacks' triggers: A frequency perspective. In *ICCV*.

Zhao, S.; Ma, X.; Zheng, X.; Bailey, J.; Chen, J.; and Jiang, Y.-G. 2020. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14443–14452.

Zheng, R.; Tang, R.; Li, J.; and Liu, L. 2022a. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision*, 175–191. Springer.

Zheng, R.; Tang, R.; Li, J.; and Liu, L. 2022b. Pre-activation Distributions Expose Backdoor Neurons. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Zhu, M.; Wei, S.; Shen, L.; Fan, Y.; and Wu, B. 2023. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4466–4477.