CL-Guard: Defending DNNs Against Backdoors via Fine-Grained Neuron Analysis and Collaborative Dual-Network Learning

Jie Xiao[1, 2], Yuhao Huang[1, 2*], Yanjiao Gao[1, 2], Aizhu Liu[1, 2], Zhezhao Yang[1, 2], Xinyue Yu[1, 2], Qianwei Zhou[1], Fan Terry Zhang [3*]

[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[2]Zhejiang Key Laboratory of Visual Information Intelligent Processing, Hangzhou 310023, China
[3]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
{xiaojiexqj, hyh077, gyj_8023, aizhul, zhezhaoyang, xinyue_yu717}@foxmail.com, zqw@zjut.edu.cn, fanzhang@zju.edu.cn

Abstract

Backdoor attacks on deep neural networks (DNNs) have garnered significant attention, particularly in edge computing applications. Given the complexity and opacity of DNNs, defending against backdoor attacks remains a formidable challenge. To address this, we propose CL-Guard, a dual-network-based defense framework designed to effectively eliminate potential backdoors in models. First, it leverages an inter-layer backpropagation algorithm to quantify each neuron's contribution to model prediction. Next, it constructs a critical neuron set through a recursive hierarchical partitioning method and an adaptive search strategy, identifying neurons critical to model prediction while minimizing the inclusion of backdoor-related neurons. Then, we perform sparse training on the non-critical neuron set, effectively strengthening the weights of critical neurons while disrupting the association between trigger features and backdoor-related neurons. Finally, we design a dual-network architecture that incorporates a fine-grained gradient backpropagation mechanism and dynamic collaborative learning, enabling the model to retain its original accuracy while preventing backdoor reactivation. The experimental results indicate that CL-Guard achieves an average Security Effectiveness Index (SEI) of approximately 95.42%, representing a 21.23% improvement over the state-of-the-art FT-SAM method.

Extended version and code —
https://github.com/huangyuhao77/CL-Guard

Introduction

Deep neural networks (DNNs) are widely used in safety-critical applications such as autonomous driving (Badjie, Cecílio, and Casimiro 2024), medical imaging (Liu et al. 2025a), and IoT edge computing (Muppasani et al. 2023). However, training DNNs often requires extensive datasets and significant computational resources, leading many developers to depend on third-party providers. This reliance introduces security vulnerabilities. For instance, a backdoor in an autonomous driving system could cause a misinterpretation of a 'stop' sign as a 'speed limit' sign, potentially leading to traffic accidents. In contrast to adversarial attacks (Liu et al. 2025b) or data reconstruction attacks (Qiu et al. 2024b), which manipulate inputs at inference time, backdoor attacks typically embed malicious behavior into the model in advance by either poisoning the training dataset or directly modifying the model. These compromised models maintain high accuracy on clean inputs (i.e., inputs without triggers) while misclassifying inputs embedded with specific triggers to an attacker-designated target class. The stealth of such attacks is further amplified by the fact that triggers are often imperceptible to the human eye, making it exceptionally challenging for users to trace the source of misclassification and discover the existence of backdoors.

To tackle this challenge, a range of backdoor defense strategies have been developed, broadly classified into pre-deployment and post-deployment approaches based on the stage of intervention. Pre-deployment defenses aim to eliminate backdoors before the model deployment, such as trigger features removal via fine-tuning (Liu, Dolan-Gavitt, and Garg 2018), backdoor disruption via weight pruning (Li et al. 2021b), and backdoor detection through weight correction (Zhu et al. 2023). In contrast, post-deployment defenses focus on identifying and purifying poisoned inputs at runtime, such as applying anomaly scoring to identify poisoned samples (Zhu et al. 2024) and disrupting triggers embedded in incoming inputs to prevent backdoor activation (Sun et al. 2023). While post-deployment defenses are effective, they tend to incur higher operational costs compared to pre-deployment methods. This study thus focuses on pre-deployment defense techniques to reduce both the incidence and diversity of backdoors, minimizing long-term maintenance costs and ensuring model security and reliability.

To address backdoor attacks more effectively, we revisits a key characteristic of such attacks, that is, in DNN models with backdoors, there often exists a stable and latent correlation between trigger features and the associated compromised neurons. This correlation can be viewed as a state of equilibrium in which these neurons remain low activated to normal inputs and highly sensitive to relevant trigger features. Disrupting this equilibrium can help to significantly mitigate the adverse effects of backdoor triggers. However, a critical challenge lies in breaking this correlation without compromising the model's predictive accuracy. To tackle this, we conducted a study based on the model's critical neurons. On the one hand, we propose a multi-dimensional neu-

*Corresponding Author.

ron partitioning method that combines an inter-layer backpropagation algorithm with a recursive hierarchical partitioning strategy to accurately extract a set of critical neurons. These neurons play a central role in the prediction process and preserving them enables the model to retain its original predictive performance. On the other hand, we apply fine-grained gradient constraints together with sparse regularization to selectively suppress non-critical neurons most strongly coupled with trigger features, severing the dependency between trigger features and backdoor-related neurons and effectively neutralizing the backdoor.

The main contributions of our article are as follows:

- We propose a gradient-based critical neuron selection method that builds a fine-grained hierarchical representation of neuron importance. By combining recursive partitioning with adaptive search, it enables precise and efficient identification of critical neurons, offering improved accuracy and adaptability over traditional coarse-grained methods.
- We propose a dual-network collaborative learning mechanism to defend backdoor attacks in DNNs, leveraging coordinated training to disrupt trigger-neuron correlations. Unlike traditional single-network approaches, this method employs two networks to dynamically refine neuron weights, achieving superior precision in isolating critical neurons and suppressing backdoor-related ones. Its collaborative framework ensures robust, efficient backdoor elimination while preserving the model's original accuracy.
- Experiments across various models, datasets, and attack types demonstrate that CL-Guard consistently outperforms existing defenses. The dual-network design offers strong scalability: S-Net adjusts sparsity to balance accuracy and defense strength, while A-Net complements neuron refinement dynamically for adaptive backdoor mitigation.

Related Works

Backdoor Attacks. Backdoor attacks can be broadly classified into pixel-space and feature-space attacks based on the type of injected malicious trigger. Pixel-space attacks modify the pixels of an image, and the triggers typically fall into three categories: local, global, and invisible. Local triggers involve small patches, such as a few pixels in a specific corner of the image (Gu, Dolan-Gavitt, and Garg 2017). Global triggers, which cover a larger area of the image, are often embedded into the background to evade detection (Chen et al. 2017). Invisible triggers, inspired by adversarial examples, apply subtle perturbations to the image, making detection more difficult (Nguyen and Tran 2021). Feature-space attacks embed triggers in the feature space, often using benign semantic features. For example, (Liu et al. 2020) uses natural semantic features unrelated to the original task to induce misclassification, while (Lin et al. 2020) activates backdoors through combinations of objects within the image. Another type of attack, sample-specific backdoor attacks, involves unique trigger patterns for each poisoned sample. (Nguyen and Tran 2020) introduces an input-aware attack with a non-reusable trigger for each input, and (Li et al. 2021a) employs invisible triggers via DNN-based image steganography. Adaptive attacks (Peng et al. 2024) are capable of bypassing most current defense mechanisms, presenting a significant security threat (Qiu et al. 2024a; Liu, Dolan-Gavitt, and Garg 2018).

Backdoor Defense. To counter the aforementioned backdoor attacks before model deployment, existing defenses are mainly categorized into pruning-based and fine-tuning-based methods. Pruning-based approaches typically remove inactive neurons identified using clean samples and then fine-tune the model to recover accuracy. For instance, (Liu, Dolan-Gavitt, and Garg 2018) prunes neurons less activated by clean data, while (Wu and Wang 2021) uses adversarial perturbations to expose and prune backdoor-related neurons. However, these methods struggle with complex attacks that affect widespread neurons, and the limited availability of clean data hinders full recovery. Fine-tuning-based defenses (Sha et al. 2022; Zeng et al. 2022) aim to overwrite backdoor behaviors. For example, (Li et al. 2021b) employs knowledge distillation by fine-tuning the original model into a partially purified teacher, then training a student model to reduce attention discrepancies. (Zhu et al. 2023) enhances this process using sharpness-aware minimization (SAM) to better suppress backdoor neurons. Although both approaches are effective, their performance instability limits their applicability in diverse scenarios. Backdoor defense still face two pressing challenges. First, there is a conflict between high coverage of various backdoor attack techniques and computational efficiency, especially under resource-constrained conditions. Second, a trade-off exists between the attack success rate (ASR) and accuracy (ACC). An effective solution to address these issues is still needed.

CL-Guard Method

Fig. 1 illustrates the framework of the proposed backdoor elimination method for DNNs.

Neuron Grading and Sparse Training

Analysis shows that a set of high-contribution neurons can retain most of the model's original performance. Concurrently, backdoor attacks typically exploit non-critical neurons with lesser task contributions, creating a hidden trigger-activated path without disrupting accuracy. Thus, fine-tuning a proper subset of these non-critical neurons can remove the backdoor while keeping normal behavior. However, accurately identifying critical neurons and quantifying their predictive contribution remains a significant challenge. Traditional methods typically focus on individual or combined neuron roles (Cao et al. 2025), whereas this paper defines critical neurons relative to backdoor neurons. This section aims to build a critical neuron set key to prediction, minimizes backdoor neuron inclusion, and then applies sparse training to non-critical neurons to enhance critical neurons' influence and break trigger-neuronlinks.

Studies on neuron interpretability show that a neuron's importance is proportional to its output contribution (Xuanyuan et al. 2023). Building on this, this paper employs

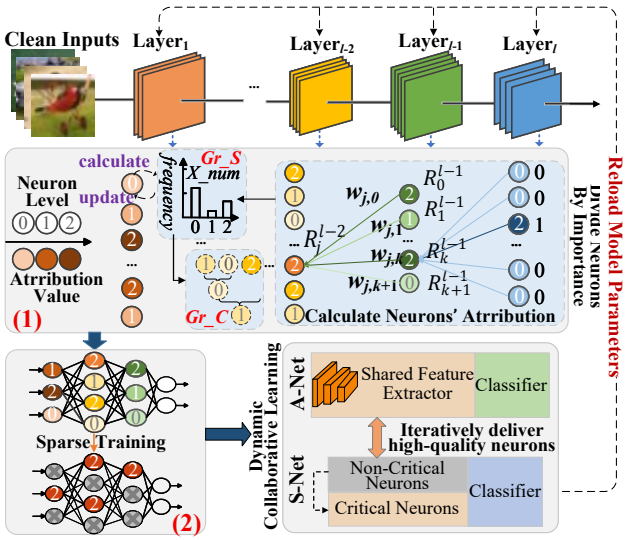

Figure 1: Flowchart of CL-Guard: (1) compute neuron contribution and constructs critical neuron set; (2) apply sparse training to suppress backdoor features; (3) refine the model through dual-network collaborative learning.

a relevance score assignment rule based on the Layer-wise Relevance Propagation (LRP) algorithm to quantify a neuron's contribution to the model output (as shown in Eq. (1)). However, selecting only high-score neurons cannot guarantee a backdoor-free set, since some backdoor neurons may also rank high for target classes, misleading selection. Given that neurons at the same position often exhibit distinct importance across different samples and class tasks, we propose a recursive hierarchical partitioning strategy based on neuron contribution computed from a set of clean samples. This strategy iteratively ranks neuron importance at the sample, class, and model levels. At the sample level, where neurons are classified into three categories according to their contribution: important (Gr = 2), secondary important (Gr = 1), and non-important (Gr = 0), as shown in Eq. (2). At the class level, we calculate each neuron's activation frequency across samples of each class and update their initial rankings, as shown in Eq. (3). Notably, neurons with higher activation are usually more important in clean models, but in backdoored models, backdoor neurons often show the highest activation in the target class (Nguyen et al. 2025). To avoid misclassification, we add a model-level ranking adjustment combining contribution scores with cross-class activation, forming a hierarchical ranking, as shown in Eq. (4).

However, an overly large critical neuron set may keep the backdoor effect active, while too few critical neurons may break weight connections and degrade accuracy. Prior work (Sun et al. 2017) shows that sparsely pruned models perform best at a pruning ratio of 0.6. Nevertheless, the skewed weight distribution, characterized by a high peak and long tail in each layer of a model, results in only a small proportion of neurons being identified as critical. This structural sparsity means that the number of important neurons (Gr = 2) obtained according to Eqs. (2)-(4) typically constitutes approximately 30% of the total number of neurons in a model. Apparently, including only the important neurons in the critical neuron set could lead to significant accuracy loss. Thus, we include some secondary neurons (Gr = 1) by randomly selecting part of them to expand the critical set, setting the final ratio of critical to total neurons to 0.6 for a balance between performance and defense. To extract a critical neuron set of appropriate size, we design an adaptive search mechanism combined with a recursive hierarchical partitioning strategy. Specifically, we initialize the critical neuron set by randomly sampling from the secondary group (Gr = 1) and combining it with the important group (Gr = 2). The model's accuracy is then evaluated using only the current critical set to assess whether the current set preserves the model predictive ability as the final key neuron set. Through iterative sampling and evaluation, the search refines the neuron set and identifies the neuron combination that best preserves accuracy under a 60% sparsity constraint as the final critical set.

$$R_j^{l-1} = \sum_k \frac{w_{j,k} \times \left(a_j^{l-1} - \tilde{a}_j^{l-1}\right)}{\sum_{0,j} w_{j,k} \times a_j^{l-1}} \cdot R_k^l. \tag{1}$$

$$Gr\_S(n_j^l) = \begin{cases} 0, R_j^l \leq Tp[Nv^l \times 2\partial]; \\ 1, Tp[Nv^l \times \partial] < R_j^l < Tp[Nv^l \times 2\partial]; \\ 2, R_j^l \geq Tp[Nv^l \times \partial]. \end{cases} \tag{2}$$

$$Gr_C(n_j^l) = \begin{cases} 0, & AF_0(n_j^l) \geq \frac{x_{\text{num}}}{2}; \\ 1, & \begin{array}{l} AF_1(n_j^l) \geq AF_2(n_j^l) \text{ and} \\ AF_0(n_j^l) < \frac{x_{\text{num}}}{2}; \end{array} \\ 2, & \text{else.} \end{cases} \tag{3}$$

$$Gr\_M(n_j^l) = \cup_c^M Gr\_C(n_j^l). \tag{4}$$

In Eq. (1), k and j denote neuron indices in layers l and l-1, respectively. $w_{j,k}$ is the weight connecting neurons k and j, a_j^{l-1} is the activation of neuron j in layer l-1, and $\tilde{a}_j^{l-1}$ is the interference component in a_j^{l-1}, which is estimated using the dual-component linear signal estimator from (Kindermans et al. 2018). R_j^{l-1} represents the contribution score of neuron j. In Eq. (2), $Gr_S(n_j^l)$ indicates the sample-level grade of neuron n_j^l. TP is the sequence of contribution scores for all neurons in layer l, sorted in descending order as $TP[1] \geq TP[2] \geq \cdots \geq TP[Nv^l]$, where Nv^l is the number of neurons in layer l, and $TP[Nv^l]$ is the contribution score of the Nv^l-th neuron. $TP[Nv^l \times \partial]$ denotes the contribution score at index ∂. As neuron activations follow a Pareto-like relevance distribution in which roughly 20–30% of neurons dominate, ∂ is typically set to $1/4$ to ensure a balanced threshold between key and secondary neurons. In Eq. (3), $Gr\_C(n_j^l)$ represents the class-level grade of neuron n_j^l, with AF_0, AF_1, and AF_2 indicating the frequencies of neurons at the same position with grades 0, 1, and 2 across samples, respectively, and x_num is the number of samples in class c of sample set X_C. In Eq. (4), $Gr\_M(n_j^l)$ denotes

the model-level grade of neuron n_j^l, determined as the final grade. M is the total number of classes, and U_c^{M} follows predefined rules: identical grades are retained, while differing grades take the higher of the two.

Analysis reveals that critical neurons consistently influence model decisions across classes, while backdoor neurons only show high contribution with poisoned samples. Thus, it can be shown that the critical neuron set selected by the above strategy typically contains only a minimal number of potential backdoor neurons. Exploiting this characteristic, we can selectively mask non-critical neurons and conduct sparse training using clean samples to strengthen the influence of critical neurons during inference, thereby weakening the role of non-critical neurons (including backdoor neurons) and disrupting the backdoor. To suppress non-critical activations and enhance critical neuron impact, we design a structured masking strategy that selectively disables non-critical neurons identified by our partitioning mechanism. This masking mechanism operates in a block-wise manner across feature maps, enabling it to simultaneously attenuate the weights of non-critical neurons and disrupt potential backdoor activation pathways. Specially, we maintain a masking matrix MM_l in every layer l of the model, where the values corresponding to critical neurons are set to 1, and all other positions are set to 0. The dimensions of the matrix MM_l are aligned with the feature map dimensions at the l-th layer. During training, we perform an element-wise multiplication between the output of the l-th layer ($Output_l$) and the corresponding masking matrix (MM_l), ensuring that only the critical neurons contribute. This method facilitates the mitigation of the backdoor effect while preserving the model's predictive ability as much as possible, by simultaneously strengthening the weights of critical neurons and suppressing those of non-critical ones.

Dynamic Collaborative Backdoor Eliminating

Existing studies show a linear link between model performance and the log of training data size, that is, under a certain model capacity, large-scale training data is particularly important for representation learning. However, for low-cost backdoor defense, defenders often have only limited data, which is also the key reason why the sparse training proposed in the previous section cannot meet the final model performance requirements. Moreover, simply strengthening critical neuron weights cannot fully remove backdoor effects, while masking most non-critical neurons often harms accuracy. To effectively address performance degradation problem caused by sparse training and further mitigate the backdoor effect, this paper proposes a dynamic collaborative learning method based on a dual-network model, as shown in Fig. 2. We define the critical neuron set and its associated weights as a sparse network (S-Net), represented as $W_s = \mathrm{Mask} \odot W$, and treat the original model's feature extractor weights as an auxiliary network (A-Net), represented as $W = (1 - \mathrm{Mask}) \odot W + \mathrm{Mask} \odot W$. Both S-Net and A-Net share the same feature extractor weights, but they employ separate classification heads: S-Net's classifier is updated with weights from sparse training, while A-Net's classifier uses the weights from the original model. This design is motivated by two core reasons: (1) A-Net calculates the gradient of the complementary weights for non-critical neurons through auxiliary classifiers, which helps select high-performing neurons for training. This allows S-Net to focus on extracting features from clean samples and reduces the risk of backdoor effects recurring; (2) S-Net and A-Net have independent classifiers, utilizing knowledge distillation to facilitate improved mutual calibration and enhance information sharing. We combine knowledge distillation (KD) loss with cross-entropy loss, using each network's output as soft targets for each Kullback-Leibler (KL) divergence term $KL(p||q;\mathcal{T})$, where $\mathcal{T}$ is the temperature value (see Eq. (8)), to effectively enhance the predictive capability of S-Net.

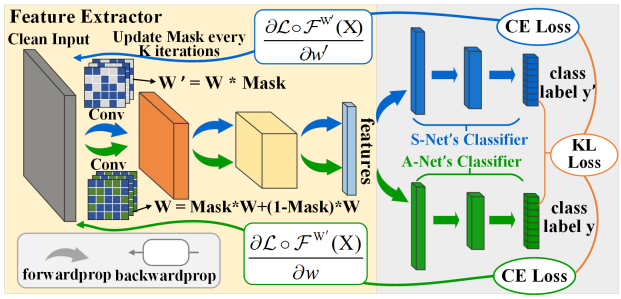

Figure 2: Dynamic Collective Learning Procedure.

$$C_p = C_f + (C_t - C_f)\left(1 - \frac{pi - pi_0}{Ep}\right)^3. \tag{5}$$

$$\mathrm{Mask}_{\mathrm{K}}^{i,j} = \begin{cases} 1, & w^{i,j} > \left[\lceil C_p \cdot \mathrm{len}(\mathrm{sorted\_}W)\rceil\right]; \\ 0, & others. \end{cases} \tag{6}$$

$$\mathrm{W} \leftarrow \mathrm{W} - \eta\{\mathrm{Mask} \odot \nabla_W \mathcal{L} + (1 - \mathrm{Mask}) \odot \nabla_{W'} \mathcal{L}\},$$
$$\nabla_{W'} \mathcal{L} \triangleq \frac{\partial \mathcal{L} \circ \mathcal{F}^{\mathrm{W}'}(\mathrm{X})}{\partial w'}, \quad \nabla_W \mathcal{L} \triangleq \frac{\partial \mathcal{L} \circ \mathcal{F}^{W}(\mathrm{X})}{\partial w}. \tag{7}$$

$$\mathcal{L} \circ \mathcal{F}(\mathbf{X}) = \mathcal{L}_{ce} \circ \mathcal{F}^{\mathbf{W}'}(\mathbf{X}) + \mathcal{L}_{ce} \circ \mathcal{F}^{\mathbf{W}}(\mathbf{X}) + \mathcal{T}^2 \cdot \mathrm{KL}\big(\mathcal{F}^{\mathbf{W}}(\mathbf{X}) \| \mathcal{F}^{\mathbf{W}'}(\mathbf{X}); \mathcal{T}\big). \tag{8}$$

Algorithm 1 presents the detailed steps of collaborative backdoor mitigation procedure. Here, F'_θ denotes the model after sparse training, and F'_θ represents the model post dual-network collaborative learning. Mask is a binary matrix matching the size of F'_θ's weights. Mask coverage m_c is the ratio of elements with value 1 to the total elements. C_p indicates mask coverage at epoch pi, where $pi \in \{pi_0, \ldots, Ep\}$, and Ep is the total number of epochs. The initial mask coverage C_f is the ratio of weights connected to critical neurons to all model weights. Mask update frequency K is proportional to the size of the clean dataset D_c; when D_c is 10% of the original dataset, K is set to 30, which aligns the update time scales of weights and masks to maintain stable optimization dynamics while allowing adaptive response to feature variations. For F'_θ, weights linked to critical neurons are among the highest in the model's weight distribution. To simplify training, we avoid manually fixing their mask values to 1. S-Net initializes its weights by selecting the top C_f-ranked weights from the model's weight distribution.

Algorithm 1: Dynamic Collaborative Backdoor Eliminating

Require: F'_θ, D_c, K
Ensure: F''_θ

Construct S-Net and A-Net;
for pi = pi_0+1 to Ep **do**
  Initialize $Iter$ = 0;
  repeat
    read mini-batch from D_c;
    while $Iter$ % K == 0 **do**
      $C_p \leftarrow$ Eq. (5), Mask $\leftarrow$ Eq. (6);
    end while
    $W \leftarrow$ Eq. (7) and $Iter$ += 1;
  until D_c has been read
end for
Reload the parameters in S-Net into the single network model, and get F''_θ.
return F''_θ

Experiment Setup

Attack Model: This paper assumes that attackers can embed backdoors in a DNN before deployment. The attacker trains a backdoored model that misclassifies poisoned samples with triggers into a target class, while maintaining normal classification performance on clean, untainted samples.

Defense Goal: We focus on a realistic scenario where the defender lacks access to the full training data and cannot retrain the model. Instead, only a small clean clean set (1%–10% of the original data) is available. The goal of the defense is to use the limited data before deployment to break potential backdoors while keeping the model's accuracy.

Attack Configurations: To validate the applicability of the CL-Guard across various backdoor attack scenarios, we considered 11 representative and advanced backdoor attack strategies spanning from pixel space to feature space: BadNets (Gu, Dolan-Gavitt, and Garg 2017), Blended (Chen et al. 2017), TrojanNet (Liu et al. 2018), SIG (Barni, Kallas, and Tondi 2019), Dynamic (Nguyen and Tran 2020), CLA (Turner, Tsipras, and Madry 2019), WaNet (Nguyen and Tran 2021), ISSBA (Li et al. 2021a), BPPA (Wang, Zhai, and Ma 2022), FBA (Zeng et al. 2021), and Refool (Liu et al. 2020). Each attack followed its original configuration for trigger design, size, and training parameters to ensure fairness. We used PreActResNet18 (Yu, Yu, and Ramalingam 2018) on CIFAR-10 (Zhang 2021) and Tiny-ImageNet (Huynh 2022) datasets, with training over 100 and 200 epochs, respectively. For GTSRB (Johner and Wassner 2019), we employed VGG-16 (Zhang 2021) with 50 epochs. All experiments adhered to standard practices, using a 10% poisoning rate (ps_r) (Wu et al. 2024), where ps_r denotes the proportion of poisoned samples in the training dataset.

Defense Configurations: We compared CL-Guard with five representative backdoor defense methods: FP (Liu, Dolan-Gavitt, and Garg 2018), NAD (Li et al. 2021b), ANP (Wu and Wang 2021), I-BAU (Zeng et al. 2022), and FT-SAM (Zhu et al. 2023). For fairness, all methods follow the default settings in BackdoorBench (Wu et al. 2024), using 10% of the benign training data. We set the learning rate to 0.01, batch size to 256, and iterations to 50 for CIFAR-10, GTSRB, and Tiny ImageNet.The experimental setup (hardware and software) is described in detail in the appendix.

Evaluation Metrics:We uses three standard metrics to evaluate the proposed backdoor defense method: attack success rate (ASR), prediction accuracy (ACC), and security effectiveness index (SEI). ASR measures the proportion of poisoned samples misclassified as the target label, while ACC reflects accuracy on clean samples. SEI, defined in Eq. (9), combines changes in ACC (ΔACC) and ASR (ΔASR) to assess defense effectiveness, with a penalty factor γ_{acc} (set to 1.0) balancing accuracy and defense performance. Higher ACC, lower ASR, and higher SEI indicate better defense. If ΔASR gain is less than the penalty, yielding a negative SEI, SEI is set to 0, indicating defense failure.

$$\text{SEI} = \frac{\Delta\text{ASR} - \gamma_{acc} \times max(\Delta\text{ACC}, 0)}{\text{ASR}_{init}} \times 100\% \quad (9)$$

Experimental Results

Performance Evaluation

Tables 1 and 2 present results on CIFAR-10 and Tiny-ImageNet, respectively, showing that CL-Guard consistently outperforms existing defenses across all evaluated backdoor attacks. It achieves notable gains in ASR, ACC, and SEI on both datasets. Due to space constraints, detailed results on the VGG architecture are presented in the appendix. CL-Guard also surpasses the best baseline on VGG, with average improvements of $<$12.26, 0.01, 11.07$>$ percentage points in ASR, ACC, and SEI.

CL-Guard begins by building a critical neuron set using a recursive hierarchical partitioning method enhanced with a targeted search strategy. These neurons are then protected via sparse training, maintaining their predictive power while disrupting backdoor-related equilibrium. A dual-network collaborative training framework further refines this defense. The collaboration between the networks aligns weight distributions, gradually eliminating residual backdoors and restoring predictive accuracy. While some backdoor neurons may remain in the critical set, sparse training weakens their connections, reducing backdoor activation. Moverover, CL-Guard demonstrates strong generalization to both clean-label and adaptive attacks. On the representative SIG clean-label poisoning, it achieves an average SEI of 92.88%, indicating its ability to suppress stealthy trigger correlations even when poisoned samples are visually identical to clean ones. Its recursive critical-neuron discovery and dual-network collaboration further ensure resilience against adaptive attacks by relying on intrinsic neuron contribution patterns rather than fixed trigger priors, effectively preventing adversarial adaptation and maintaining stable defense performance.Overall, the neuron selection mechanism and collaborative training strategy enable CL-Guard to effectively adapt to complex data distributions and scale to larger datasets, achieving robust performance across various attack scenarios, including Dynamic and FBA. In contrast, existing defense methods often suffer significant drop in ACC, ASR and SEI on larger datasets due to their limited ability to isolate or suppress backdoor-related neurons.

Types	Attacks	No defense		FP			ANP			NAD			I-BAU			FT-SAM			CL-Guard(OURS)		
		ASR	ACC	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI
Pixel Space	Badnets	93.53	91.66	0.97	91.80	98.96	86.20	58.38	0.00	0.23	90.32	95.98	0.37	87.69	94.59	1.74	91.38	97.84	0.32	91.70	99.65
	Blended	98.23	93.37	5.61	92.90	93.81	95.78	90.27	0.00	0.36	87.69	96.09	0.60	89.02	96.64	18.17	92.71	80.83	0.43	91.20	97.35
	TrojanNet	99.99	93.40	87.87	92.82	11.54	99.99	93.40	0.00	2.40	92.40	98.29	4.08	89.94	96.22	1.19	92.60	98.00	3.76	92.67	95.51
	SIG	95.63	93.64	42.16	93.27	55.53	72.52	84.41	14.51	35.93	92.61	79.33	0.01	91.41	96.69	0.20	92.98	99.10	0.50	92.53	98.32
	CLA	92.89	93.49	63.78	92.60	30.38	19.03	88.31	73.93	84.35	92.57	53.81	10.58	91.17	89.99	42.72	93.05	53.53	6.32	92.25	91.86
	Dynamic	98.92	91.56	12.60	93.72	87.36	98.23	89.87	0.00	53.61	93.33	72.65	10.03	91.68	94.44	17.82	93.44	81.98	7.67	92.60	92.25
	ISSBA	99.77	93.70	6.81	91.48	90.94	95.44	85.37	0.00	0.18	92.08	98.98	22.84	90.07	86.65	0.74	92.57	98.12	0.74	92.30	97.08
	WaNet	97.73	90.46	1.01	93.49	98.97	68.95	87.14	26.05	0.17	93.09	98.78	0.46	92.37	98.63	0.23	93.29	99.76	0.35	93.33	99.64
	BPPA	99.81	90.55	1.31	93.24	98.68	75.66	82.62	16.25	0.52	93.12	99.64	1.18	90.79	99.31	0.26	93.44	99.73	1.23	93.60	98.78
Feature Space	Refool	93.24	92.35	10.91	92.21	86.34	91.80	88.67	0.00	16.83	91.88	87.97	11.02	90.80	90.33	3.57	92.25	96.06	2.12	92.30	97.67
	FBA	99.26	93.34	21.99	92.85	76.65	98.54	92.70	0.08	83.52	92.25	57.32	2.21	90.41	97.06	1.89	92.61	97.36	2.97	92.87	94.52
Avg. on the above attacks		–	–	23.18	92.76	73.37	81.48	86.08	11.89	25.28	91.94	85.35	5.76	90.48	94.59	8.04	92.75	91.12	2.58	92.48	96.67

Table 1: Comparison with state-of-the-art mitigations on CIFAR-10 with 10% benign data on PreAct-ResNet18 (%).

Types	Attacks	No defense		FP			ANP			NAD			I-BAU			FT-SAM			CL-Guard(OURS)		
		ASR	ACC	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI	ASR	ACC	SEI
Pixel Space	Badnets	99.96	56.66	25.40	52.03	69.95	17.60	45.17	70.89	0.14	49.16	92.35	77.29	54.09	20.10	0.03	51.40	94.70	0.13	52.50	95.71
	Blended	97.43	56.43	48.82	51.61	44.94	94.32	46.84	0.00	62.71	48.73	27.73	11.25	52.28	84.19	85.00	49.94	6.09	0.69	51.20	93.92
	TrojanNet	99.99	56.50	87.80	51.97	7.66	99.92	55.65	0.00	0.51	49.69	92.67	1.67	56.02	97.84	0.56	50.69	93.62	0.20	52.65	95.95
	SIG	89.95	58.09	48.21	53.27	41.04	90.17	56.91	0.00	84.30	51.78	0.00	17.48	55.44	77.62	88.99	47.66	0.00	5.37	52.15	87.43
	Dynamic	99.80	58.28	0.29	54.64	96.06	8.27	54.45	87.87	0.68	50.86	91.88	1.97	55.81	95.55	0.30	50.26	91.66	0.29	53.65	95.07
	WaNet	89.58	57.78	9.07	52.77	84.28	75.42	57.57	15.57	0.25	46.87	87.54	3.66	57.26	95.33	0.73	48.54	88.87	0.25	52.45	93.77
	BPPA	99.97	58.45	0.47	53.64	94.71	33.46	57.33	65.40	0.06	47.46	88.94	60.38	57.54	38.69	0.06	49.76	91.24	0.17	53.65	95.02
Feature Space	Refool	99.07	56.64	60.10	51.97	34.62	87.27	52.37	7.60	52.17	49.58	40.21	39.34	54.21	57.83	35.57	49.85	57.24	5.32	52.15	90.09
	FBA	98.34	56.15	59.69	51.31	34.38	98.03	52.33	0.00	29.85	46.90	60.23	75.26	52.72	19.98	89.85	48.97	1.33	6.65	52.45	89.47
Avg. one the above attacks		–	–	37.76	52.58	56.41	67.16	53.18	27.48	25.63	49.00	64.54	32.03	55.04	65.24	33.45	49.67	58.30	2.11	52.54	92.94

Table 2: Comparison with state-of-the-art mitigations on Tiny-ImageNet with 10% benign data on PreAct-ResNet18 (%).

Efficiency Investigation

To evaluate the novelty and effectiveness of the proposed method, we conducted two sets of experiments. First, we evaluated backdoor removal effectiveness before and after applying the critical neurons selection method, and compared the dual-network cooperative strategy with traditional fine-tuning method (Sha et al. 2022). The results are shown in Fig. 3, where Sc-1 and Sc-2 represent the proposed method without critical neurons selection method (using randomly selected neurons instead) and without dual-network cooperation, respectively. Second, we analyzed the effect of varying poisoning rates of ps_r (1%, 5%, 25%, and 50%) on the performance of CL-Guard across different datasets and attack types. The results are shown in Fig. 4.

Fig. 3 shows that the CL-Guard consistently outperforms Sc-1 and Sc-2 across all metrics. It achieves higher ACC, lower ASR, and significantly improved SEI, demonstrating its robust defense capabilities. The critical neuron identification method proposed in Section **Neuron Grading and Sparse Training**, along with the dual-network collaborative training strategy presented in Section **Dynamic Collaborative Backdoor Eliminating**, not only effectively enhances the feature representation capabilities of critical neurons to safeguard model performance but also significantly reduces the backdoor effect by cutting most of the connections between backdoor neurons. By contrast, Sc-1 employs sparse training based on randomly selected neurons instead of critical neurons. While this may help the model forget the backdoor, it makes it more challenging to restore the model's predictive performance. Moreover, the random selection of neurons increases the likelihood of inadvertently selecting

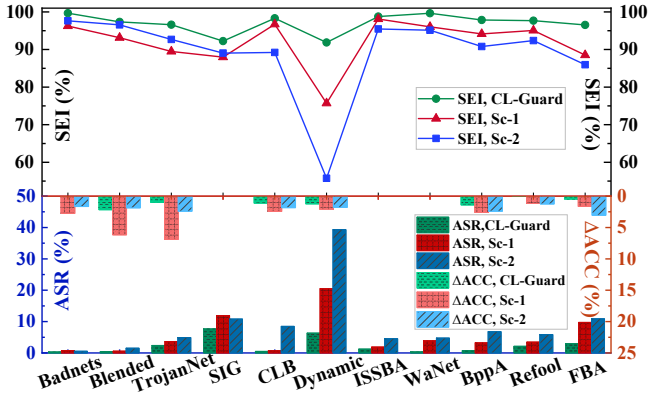

Figure 3: Effect of CL-Guard on Eliminating backdoors before and after applying the different modules.

backdoor neurons, thereby raising the probability of residual backdoor neurons within a network. Although Sc-2 also utilizes critical neurons for sparse training, its fine-tuning process differs from the proposed approach. Unlike the dual-network collaborative training strategy, which progressively selects high-quality neurons from the remaining neurons for learning, Sc-2's fine-tuning disperses the weights of critical neurons across other neurons to adapt to more clean samples. This increases the risk of reactivating backdoor neurons and thus makes Sc-2 less effective in mitigating the backdoor effect compared with the proposed method.

Fig. 4 shows that the CL-Guard performs well against both attacks on these datasets, outperforming other defenses

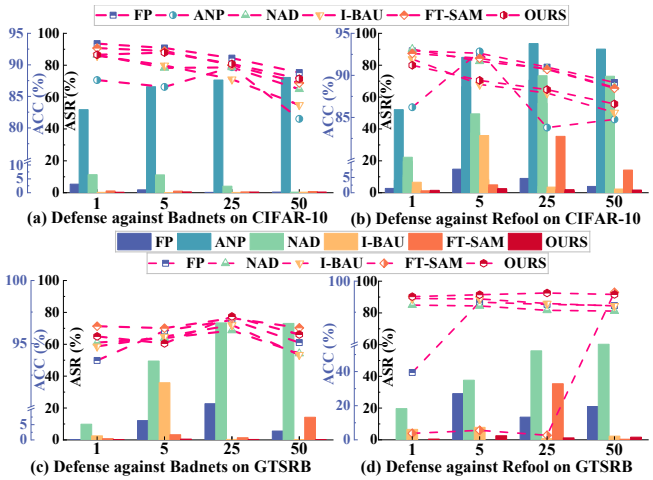

Figure 4: Comparison of defense performance of different defense methods with different ps_r (%).

on average. Relative to the best average SEI across the four scenarios, our method achieves improvements of <0.08%, 2.88%, -1.89%, 19.52%>, demonstrating its strong defensive capability. A larger poisoning rate ps_r indicates that more poisoned samples are used during model training, which clearly enhances the backdoor effect of the model while simultaneously reducing its ability to fit clean samples. Conversely, a lower ps_r typically results in fewer backdoor neurons in a model, which increases the difficulty for FP to remove backdoors by pruning dormant neurons. CL-Guard selects critical neurons solely based on contribution scores computed from clean samples. Backdoor neurons generally exhibit low scores under clean inputs and rarely rank highly regardless of the poisoning rate. Consequently, backdoor neurons are unlikely to be included in the critical neuron set under any ps_r, allowing CL-Guard to maintain robust performance across diverse scenarios.

Further Exploration

CL-Guard supports programmable model compression by dynamically adjusting the final mask coverage (m_c) in the dual-network collaboration strategy, adapting to diverse computational and storage needs across applications. Fig. 5 shows the defense effectiveness against backdoor attacks on CIFAR-10 and Tiny-ImageNet, tested with m_c values of 0.5, 0.6, 0.7, 0.8, and 1.0. Higher m_c indicates a less sparse model, with m_c = 1.0 representing an unpruned model.

Fig. 5 shows that reducing m_c typically lowers both accuracy and defense effectiveness, but the extent of this impact varies depending on the specific attack scenario. In the case of clean-label attacks such as SIG and CLA, the SEI decreases only slightly as m_c is reduced. This suggests that for simpler attack types, moderate reductions in mask coverage do not significantly affect the defense capability. In contrast, for more complex and sophisticated attacks, such as Blended, a lower m_c can enhance defense performance by reducing the likelihood of backdoor reactivation. This occurs because with fewer mask nodes in place, the backdoor's

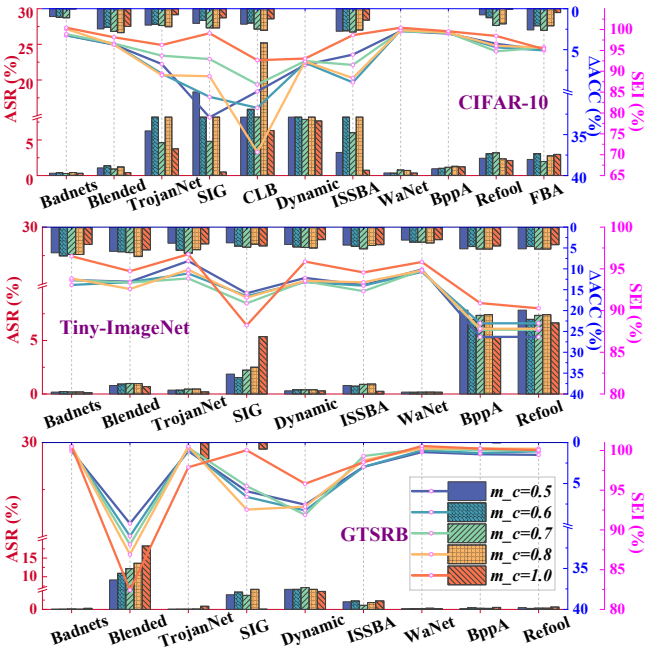

Figure 5: Effect of CL-Guard on eliminating backdoors when valuing different m_c (bars are omitted where ASR or ΔACC equals zero for certain attacks).

influence is mitigated, which results in a stronger defense. Furthermore, the collaborative training strategy constructed by proposed method benefits from model compression. By promoting a more efficient allocation of model resources and emphasizing feature learning, this strategy fosters improved defense performance, particularly against more intricate attack types. This indicates that the approach is not only effective in handling diverse attack scenarios, but also efficient in terms of model size, making it suitable for real-world applications where computational resources are often limited. Overall, adjusting m_c provides a flexible way to balance model lightweighting and defense robustness, enabling scalable protection for real-world DNNs with limit resources.

Conclusion

Effectively defending DNNs against backdoor attacks is key to securing intelligent systems. This paper presents CL-Guard, a defense that removes hidden backdoors before deployment. We propose a gradient-guided critical neuron selection with targeted sparse optimization to separate trigger signals from benign features, cutting backdoor paths while keeping model accuracy. A dual-network collaborative learning strategy with fine-grained gradient alignment further lowers the risk of reactivation by strengthening feature robustness and improving defense stability. Experiments on multiple datasets and models confirm the method's strong effectiveness and robustness. Future work will extend CL-Guard to object detection and segmentation by adapting its neuron analysis and collaborative optimization to task-specific architectures for better generality and robustness.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Zhejiang Province (No.LR24F020003), the National Natural Science Foundation of China (No.62472386), and the National Key R&D Program of China (No.2023YFB3106800).

References

Badjie, B.; Cecílio, J.; and Casimiro, A. 2024. Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review. *ACM Computing Surveys*, 57(1): 1–52.

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 2019, Taipei, Taiwan, September 22-25, 2019*, 101–105. IEEE.

Cao, T. M.; Hoang-Xuan, N.; Pham, H. H.; Nguyen, P. L.; and Thai, M. T. 2025. NeurFlow: Interpreting Neural Networks through Neuron Groups and Functional Interactions. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *CoRR*, abs/1712.05526.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *CoRR*, abs/1708.06733.

Huynh, E. 2022. Vision Transformers in 2022: An Update on Tiny ImageNet. *CoRR*, abs/2205.10660.

Johner, F. M.; and Wassner, J. 2019. Efficient Evolutionary Architecture Search for CNN Optimization on GTSRB. In Wani, M. A.; Khoshgoftaar, T. M.; Wang, D.; Wang, H.; and Seliya, N., eds., *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*, 56–61. IEEE.

Kindermans, P.; Schütt, K. T.; Alber, M.; Müller, K.; Erhan, D.; Kim, B.; and Dähne, S. 2018. Learning how to explain neural networks: PatternNet and PatternAttribution. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021a. Invisible Backdoor Attack with Sample-Specific Triggers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 16443–16452. IEEE.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021b. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Lin, J.; Xu, L.; Liu, Y.; and Zhang, X. 2020. Composite Backdoor Attack for Deep Neural Network by Mixing Existing Benign Features. In Ligatti, J.; Ou, X.; Katz, J.; and Vigna, G., eds., *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, 113–131. ACM.

Liu, C.; Cao, Y.; Zhang, Y.; Su, X.; and Zhu, H. 2025a. Perturbating, Tuning, and Collaborating: Harnessing Vision Foundation Models for Single Domain Generalization on Medical Imaging. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 5370–5378. AAAI Press.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In Bailey, M. D.; Holz, T.; Stamatogiannakis, M.; and Ioannidis, S., eds., *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings*, volume 11050 of *Lecture Notes in Computer Science*, 273–294. Springer.

Liu, X.; Ma, J.; Wang, X.; Lin, Q.; Zhang, J.; Schaefer, G.; Turkay, C.; and Fang, H. 2025b. Recoverable Facial Identity Protection via Adaptive Makeup Transfer Adversarial Attacks. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 514–522. AAAI Press.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.; Zhai, J.; Wang, W.; and Zhang, X. 2018. Trojaning Attack on Neural Networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society.

Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, volume 12355 of *Lecture Notes in Computer Science*, 182–199. Springer.

Muppasani, B.; Anand, C. J.; Appajigowda, C.; Srivastava, B.; and Johri, L. 2023. A Dataset and Baseline Approach for Identifying Usage States from Non-intrusive Power Sensing with MiDAS IoT-Based Sensors. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 15545–15550. AAAI Press.

Nguyen, D. T.; Tran, N. N.; Johnson, T. T.; and Leach, K. 2025. PBP: Post-training Backdoor Purification for Malware Classifiers. In *32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 24-28, 2025*. The Internet Society.

Nguyen, T. A.; and Tran, A. T. 2020. Input-Aware Dynamic Backdoor Attack. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nguyen, T. A.; and Tran, A. T. 2021. WaNet - Imperceptible Warping-based Backdoor Attack. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Peng, H.; Qiu, H.; Ma, H.; Wang, S.; Fu, A.; Al-Sarawi, S. F.; Abbott, D.; and Gao, Y. 2024. On Model Outsourcing Adaptive Attacks to Deep Learning Backdoor Defenses. *IEEE Trans. Inf. Forensics Secur.*, 19: 2356–2369.

Qiu, H.; Ma, H.; Zhang, Z.; Abuadbba, A.; Kang, W.; Fu, A.; and Gao, Y. 2024a. Toward a Critical Evaluation of Robustness for Deep Learning Backdoor Countermeasures. *IEEE Trans. Inf. Forensics Secur.*, 19: 455–468.

Qiu, P.; Zhang, X.; Ji, S.; Fu, C.; Yang, X.; and Wang, T. 2024b. HashVFL: Defending Against Data Reconstruction Attacks in Vertical Federated Learning. *IEEE Trans. Inf. Forensics Secur.*, 19: 3435–3450.

Sha, Z.; He, X.; Berrang, P.; Humbert, M.; and Zhang, Y. 2022. Fine-Tuning Is All You Need to Mitigate Backdoor Attacks. *CoRR*, abs/2212.09067.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 843–852. IEEE Computer Society.

Sun, T.; Pang, L.; Chen, C.; and Ling, H. 2023. Mask and Restore: Blind Backdoor Defense at Test Time with Masked Autoencoder. *CoRR*, abs/2303.15564.

Turner, A.; Tsipras, D.; and Madry, A. 2019. Label-Consistent Backdoor Attacks. *CoRR*, abs/1912.02771.

Wang, Z.; Zhai, J.; and Ma, S. 2022. BppAttack: Stealthy and Efficient Trojan Attacks against Deep Neural Networks via Image Quantization and Contrastive Adversarial Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 15054–15063. IEEE.

Wu, B.; Chen, H.; Zhang, M.; Zhu, Z.; Wei, S.; Yuan, D.; Zhu, M.; Wang, R.; Liu, L.; and Shen, C. 2024. BackdoorBench: A Comprehensive Benchmark and Analysis of Backdoor Learning. *CoRR*, abs/2407.19845.

Wu, D.; and Wang, Y. 2021. Adversarial Neuron Pruning Purifies Backdoored Deep Models. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 16913–16925.

Xuanyuan, H.; Barbiero, P.; Georgiev, D.; Magister, L. C.; and Liò, P. 2023. Global Concept-Based Interpretability for Graph Neural Networks via Neuron Analysis. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 10675–10683. AAAI Press.

Yu, X.; Yu, Z.; and Ramalingam, S. 2018. Learning Strict Identity Mappings in Deep Residual Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 4432–4440. Computer Vision Foundation / IEEE Computer Society.

Zeng, Y.; Chen, S.; Park, W.; Mao, Z.; Jin, M.; and Jia, R. 2022. Adversarial Unlearning of Backdoors via Implicit Hypergradient. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zeng, Y.; Park, W.; Mao, Z. M.; and Jia, R. 2021. Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 16453–16461. IEEE.

Zhang, X. 2021. The AlexNet, LeNet-5 and VGG NET applied to CIFAR-10. In *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 414–419. IEEE.

Zhu, L.; Ning, R.; Li, J.; Xin, C.; and Wu, H. 2024. SEER: Backdoor Detection for Vision-Language Models through Searching Target Text and Image Trigger Jointly. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 7766–7774. AAAI Press.

Zhu, M.; Wei, S.; Shen, L.; Fan, Y.; and Wu, B. 2023. Enhancing Fine-Tuning based Backdoor Defense with Sharpness-Aware Minimization. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 4443–4454. IEEE.