Logit-Margin Repulsion for Backdoor Defense

Zhiguo Yang, Dongsheng Xu, Ruizhi Zhong, Jiacheng Pi, Xingxing Huang, Wenjie Ruan*
School of Computer Science and Technology, University of Science and Technology of China
{zhiguoyang, xds2327659729, zhongruizhi, pijch, huangxingxing}@mail.ustc.edu.cn, rwjie@ustc.edu.cn

Abstract

*Backdoor attacks pose a significant threat to deep neural networks. Recent studies have shown that model compression, such as quantization and pruning, can be exploited by attackers to implant conditional backdoors. Such backdoors remain dormant in the original model but are activated after the model undergoes specific operations, making them highly stealthy and difficult to detect. Traditional defense methods struggle to counter this type of attack, while defenses specifically designed for conditional backdoors also have difficulty handling traditional backdoor attacks. To address these challenges, we propose a universal defense method, termed **L**ogit **M**argin **R**epulsion (LMR). LMR uses a small set of clean samples and combines selective cross-entropy with a logit-margin constraint to enlarge the gap between the backdoor class and benign classes. It then removes channels associated with backdoor behavior through selective pruning, thereby achieving strong backdoor purification. Extensive experiments on a variety of CNNs and Vision Transformers demonstrate that, even with an extremely limited amount of clean data (0.1%), LMR can effectively mitigate both traditional and conditional backdoor attacks. The implementation is publicly available on* https://github.com/Trusted-LLM/LMR.

1. Introduction

Deep learning has achieved remarkable progress in safety-critical tasks such as autonomous driving and face recognition [12, 39, 57, 61]. However, deep neural networks (DNNs) face serious security threats in real-world deployment and are particularly vulnerable to backdoor attacks [20–22, 38, 52, 58]. Attackers can implant backdoors during training through data poisoning and related techniques, causing the model to behave normally on clean samples while outputting a target label when a specific trigger is present [14, 46]. Meanwhile, as deep models are increasingly deployed in resource-constrained scenarios, model compression techniques have also become a new attack surface [16, 48]. Attackers can exploit parameter rounding or structural changes during model compression to implant backdoors. As shown in Fig. 1, such backdoors usually remain dormant in the original model, thereby evading traditional detection and purification defenses; however, they are activated after quantization or pruning, greatly increasing both their stealthiness and destructiveness [8, 48]. To address the threats posed by backdoor attacks, existing defense methods can be broadly divided into two categories: backdoor detection and backdoor purification. Detection-based methods aim to identify whether a model or dataset has been poisoned [1], whereas purification-based methods attempt to remove malicious behaviors from infected models [33]. Although these methods have achieved significant success in traditional backdoor attack scenarios, they face clear limitations in conditional backdoor settings: conditional backdoor models usually behave similarly to benign models in their original state, making it difficult

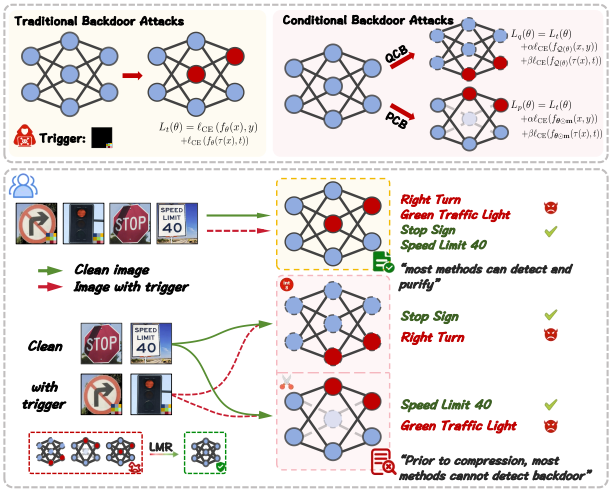

Figure 1. Schematic diagram of traditional and conditional backdoor attacks (e.g, ***Q***uantization-***C***onditioned ***B***ackdoor / ***P***runing-***C***onditioned ***B***ackdoor). The proposed LMR method defends against both types effectively.

*Corresponding author.

for traditional detection and purification methods to capture the anomalous features that become dominant only after the model undergoes specific operations, and thus preventing them from effectively suppressing backdoor behaviors. To address this issue, prior work has analyzed the implantation mechanisms of quantization-conditioned backdoors and proposed corresponding defense methods [26, 27]; however, such methods are difficult to generalize to traditional backdoor attacks. Therefore, there is still no universal defense capable of resisting both conditional and traditional backdoors.

The primary effect of a backdoor attack is the abnormal elevation of the target-class logit after a trigger is applied: when the trigger makes the logit corresponding to the backdoor class the largest one, the model prediction is flipped. For conditional backdoors, specific model operations can shift logits that are originally close to normal states, thereby activating the backdoor [17, 54]. Based on this basic phenomenon, we find that even without prior knowledge of the backdoor, simply enlarging the logit margin between the backdoor class and its strongest competing class on clean samples can significantly reduce the attack success rate (ASR), by making the shifts induced by triggers or conditional operations insufficient to alter the top-1 prediction.

We propose Logit-Margin Repulsion, a universal backdoor defense framework. LMR first accurately identifies the backdoor class through unlearning [33, 51], and then purifies the backdoor through a two-stage process. In **Stage I**, LMR uses a small amount of clean data and combines selective cross-entropy to weaken the supervision on the backdoor class, while imposing a logit penalty on all non-backdoor samples to maximize the logit margin between the backdoor class and benign classes. In **Stage II**, LMR performs selective pruning according to the ℓ_1 variation of channels associated with the backdoor class before and after Stage I, removing the channels with the most significant changes to reduce the risk of backdoor reactivation during subsequent fine-tuning. We conduct comprehensive evaluations of LMR on CNNs and Vision Transformers (ViTs) [9] using CIFAR-10 [24], Tiny-ImageNet, and ImageNet [6], covering nine representative traditional backdoor attacks and three conditional backdoor attacks, including two quantization-conditioned backdoors and one pruning-conditioned backdoor. Cross-architecture and cross-dataset results show that, for traditional backdoor attacks, LMR significantly reduces ASR while maintaining or even slightly improving accuracy. For more complex and stealthier conditional backdoor attacks, LMR still maintains stable and robust purification performance. Our main **contributions** are as follows:

- We demonstrate that enlarging the logit margin between non-backdoor classes and the backdoor class on clean samples can significantly weaken the backdoor effect.
- We propose LMR, a universal backdoor purification algorithm. Extensive experiments show that LMR requires only a small amount of clean data (even $<0.1\%$) to effectively suppress backdoor behaviors and prevent their reactivation during subsequent training.
- Extensive experiments show that LMR not only exhibits robust purification performance against traditional backdoor attacks, but is also effective in more stealthy conditional backdoor scenarios.

2. Related Work

Since early studies revealed the vulnerabilities of backdoors [4, 35, 38], related research has developed rapidly. Backdoor attacks can be broadly categorized into two types: **traditional backdoors** and **conditional backdoors**.

2.1. Traditional Backdoor

Traditional Backdoor Attack: Traditional backdoor attacks can be broadly categorized into two types: **input-space** and **feature-space**. Input-space attacks typically implant backdoors by tampering with training samples, i.e., embedding triggers into samples from non-target classes and modifying their labels to the target label. Trigger patterns take various forms, including black-and-white patches [14], random noise, sinusoidal stripes [1], blended backgrounds [4], and sample-specific trigger patterns [30]. Feature-space attacks, in contrast, implant backdoors by optimizing objectives over intermediate representations [5, 62], and are usually more stealthy.

Traditional Backdoor Defense: Traditional backdoor defenses can be broadly divided into two categories: **backdoor detection** and **backdoor purification**. Detection-based methods mainly focus on identifying whether the model [1, 3, 23, 53] or the data [2, 18, 43, 49] has been contaminated. Some methods can even reverse-engineer the trigger [47, 51, 55]. Purification-based methods typically remove backdoors from the model by fine-tuning [34, 65] or pruning [31, 33, 56, 63] to eliminate neurons or channels associated with the backdoor [32].

2.2. Conditional Backdoor

Conditional Backdoor Attack: Conditional backdoor attacks include **quantization-conditioned backdoors** [16, 37] and **pruning-conditioned backdoors**. Quantization converts full-precision models into low-bit representations through parameter rounding, substantially reducing storage and computation costs [13, 64]. Quantization-conditioned backdoors exploit this process to implant stealthy malicious behaviors [8, 10, 11, 42]. In contrast, pruning-conditioned backdoors [48] maliciously leverage the standard pruning

pipeline, encoding backdoor behaviors into the subnetwork that remains after pruning. As shown in Eqs. (1) and (2), conditional backdoors remain dormant in the original model but are activated after specific operations.

$$f(x) = f(\tau(x)) = y,\ f_Q(x) = y,\ f_Q(\tau(x)) = t. \quad (1)$$

$$f(x) = f(\tau(x)) = y,\ f_{\boldsymbol{\theta}\odot\mathbf{m}}(x) = y,\ f_{\boldsymbol{\theta}\odot\mathbf{m}}(\tau(x)) = t. \quad (2)$$

Here, f denotes the full-precision model, f_Q denotes the corresponding k-bit quantized model, $\tau(\cdot)$ denotes the trigger injection operation, θ denotes the model parameters, $m \in \{0, 1\}$ denotes an element-wise pruning mask, $\odot$ denotes the Hadamard product, and t denotes the attacker's target class.

Table 1. Coverage of existing defenses against traditional backdoor attacks (TBA) and conditional backdoor attacks (CBA), including quantization-conditioned (CBA-QCB) and pruning-conditioned (CBA-PCB) attacks.

DEFENSE	TBA	CBA-QCB	CBA-PCB
FP [34]	✓	✗	✗
I-BAU [60]	✓	✗	✗
RNP [33]	✓	✗	✗
MNP [28]	✓	✗	✗
EFRAP [27]	✗	✓	✗
LACPDA [26]	✗	✓	✗
LMR (Ours)	✓	✓	✓

Conditional Backdoor Defense: As shown in Tab. 1, existing traditional defense methods still exhibit limitations when addressing conditional backdoor attacks [16]. For quantization-conditioned backdoors, *Li* guides rounding direction adjustments through error-driven optimization [27]. *Li* achieves backdoor purification by aligning the quantized model with the full-precision model [26]. However, such specialized defense methods have limited effectiveness against traditional backdoors and are also difficult to apply to pruning-conditioned backdoor attack scenarios.

3. Methodology

3.1. Threat Model

We consider two threat scenarios: (i) traditional backdoor attacks and (ii) conditional backdoor attacks. The defender has access to the model logits, with the objective of effectively removing the backdoor while maintaining the model's normal task performance. It is assumed that the defender possesses only a very small amount of clean data (e.g., about 1% of the original training set).

3.2. Logit Margin Repulsion

Fig. 2 illustrates the overall pipeline of LMR. In the **anti-learned** stage, the model's classification accuracy on clean samples is reduced to near-random levels, thereby localizing the potential backdoor class. The model then enters a two-phase purification process. In **Phase-1**, a dedicated loss function is used to enlarge the logits margin between the backdoor class and the correct class, thereby suppressing the backdoor behavior. At the same time, selective cross-entropy and conditional margin constraints are introduced to maintain stable discrimination on non-backdoor classes. In **Phase-2**, channels highly related to the backdoor class are screened and pruned, while a lightweight fine-tuning step is applied to improve the benign accuracy of the target class. Meanwhile, we compare LMR with two representative methods. FP removes channels or neurons based on weight magnitude or activation strength, but is not effective enough at removing more stealthy backdoors and is also prone to over-pruning. RNP adopts an asymmetric unlearning–recovery procedure to expose and prune backdoor channels, but under more stealthy attacks, it may also damage normal channels. In contrast, LMR achieves more precise backdoor purification while preserving the representations of non-backdoor classes as much as possible.

Backdoor Class Estimation: We perform anti-learned on a small batch of clean samples by maximizing the cross-entropy loss:

$$\mathcal{L}(x, y; \theta) = -\frac{1}{m}\sum_{i=1}^{m} \mathrm{CE}(f_\theta(x_i),\ y_i)\,. \quad (3)$$

This process significantly suppresses the activation of normal neurons while leaving backdoor-related neurons largely unaffected, thereby revealing potential backdoor biases. Subsequently, based on the parameters θ' obtained after anti-learned, we compute the softmax posterior on the clean batch. For each class, we calculate the mean of the log probabilities over the batch and identify the class with the highest mean as the backdoor class:

$$s(c) = \frac{1}{m}\sum_{i=1}^{m} \log p_{\theta'}(y = c \mid x_i), \quad \hat{y}_t = \arg\max_c\ s(c). \quad (4)$$

We set the backdoor class index to $c \coloneqq \hat{y}_t$. In Appendix B, we applied this method to backdoor localization tests across multiple models and datasets, achieving 100% localization accuracy in all cases.

Two-Stage Purification Framework: We systematically suppress the model's response to the backdoor class on the clean distribution while maintaining stable decision boundaries for non-backdoor classes in Phase-1. In Phase-2, a light fine-tuning is performed using a small set of clean samples to further restore the model's recognition accuracy.

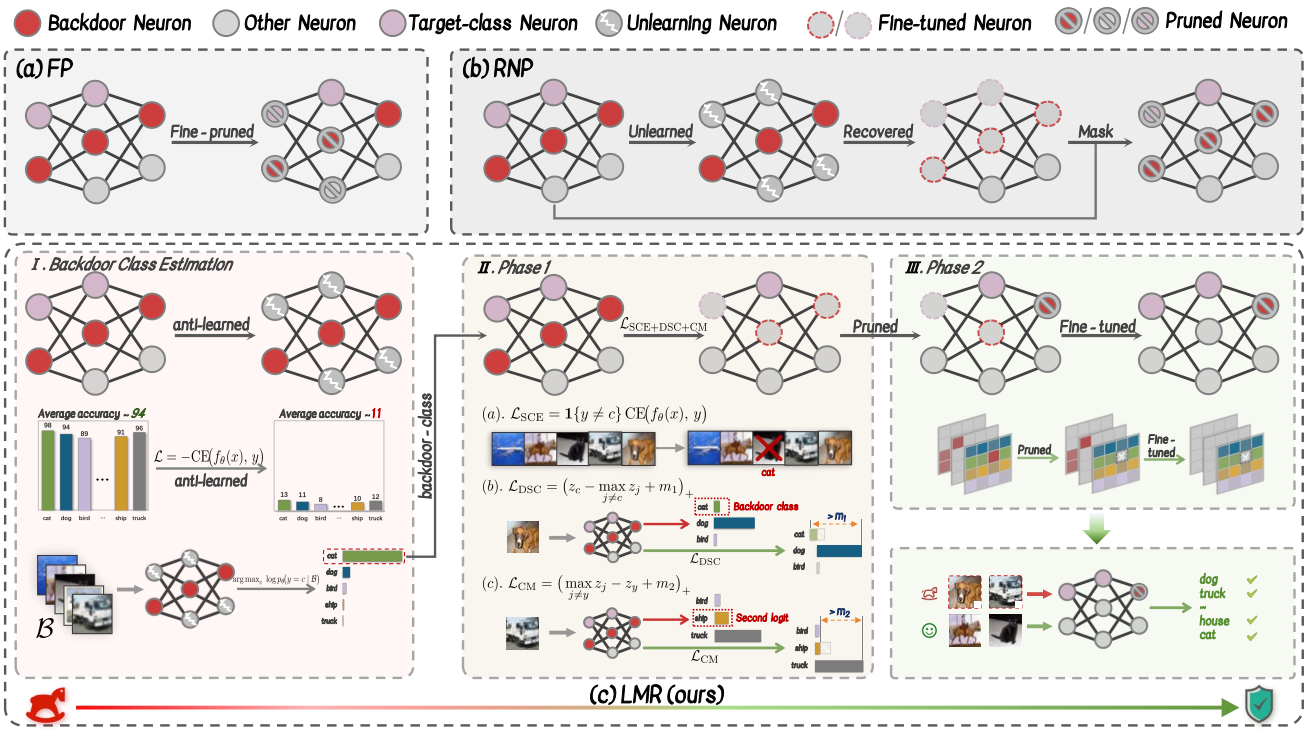

Figure 2. A comparative overview of our proposed LMR method against two existing backdoor removal approaches (FP and RNP). LMR compresses the decision space of the backdoor class by applying $\mathcal{L}_{\mathrm{SCE}}$, $\mathcal{L}_{\mathrm{DSC}}$ and $\mathcal{L}_{\mathrm{CM}}$.

Phase 1: Logit-Geometric Purification: Extensive studies have shown that when an input contains a trigger [17, 54], the *logits* corresponding to the backdoor class *abnormally increase*, strongly suppressing the responses of other classes. In geometric terms, this manifests as an abnormal expansion of the backdoor class's decision region. Based on this observation, we design three loss functions to reshape the backdoor decision boundary on clean data, thereby purifying the latent backdoor within the model.

(I). Selective Cross-Entropy (SCE): To avoid suppressing the accuracy of other classes during the purification phase or unintentionally reinforcing backdoor-related representations, we temporarily disable the CE weight (set to 0) for samples with label y = c.

$$\mathcal{L}_{\mathrm{SCE}}(x, y; \theta) = \mathbf{1}\{y \neq c\}\, \mathrm{CE}\big(f_\theta(x),\, y\big). \tag{5}$$

(II). Directed Suppression of Backdoor-Class Logits (DSC): For all clean samples with labels $y \neq c$, we require the *logit margin* against the backdoor class to exceed a positive margin m_1:

$$\mathcal{L}_{\mathrm{DSC}}(x, y; \theta) = \big(z_c - \max_{j \neq c} z_j + m_1\big)_+ \mathbf{1}\{y \neq c\}, m_1 > 0. \tag{6}$$

Although imposed only on clean data, this constraint markedly suppresses the model's response to the backdoor class on triggered samples at test time, geometrically contracting the backdoor decision region. While it may slightly affect the accuracy of benign samples from class c, this can be corrected in subsequent stages. Importantly, the loss does not assume that clean samples naturally have high backdoor logits—the constraint is actively constructed on the clean distribution, making it applicable to stealthier backdoors. Consequently, any perturbation that pushes a sample into class c must cross a larger margin. Appendices C and D confirm that a logit margin $> m_1$ on clean samples significantly mitigates backdoors.

(III). Conditional Margin (CM): DSC may introduce error shifts to non-target class, causing boundary jitter. To enhance stability, we add a conditionally activated regularizer that penalizes a sample only when the true-class response does not lead its closest competitor (i.e., ambiguous / borderline cases). For confident samples, the penalty is zero.

$$\mathcal{L}_{\mathrm{CM}}(x, y; \theta) = \big(\max_{j \neq y} z_j - z_y + m_2\big)_+, m_2 > 0. \tag{7}$$

Phase-1 Loss Function:

$$\mathcal{L}_{\mathrm{P1}}(\theta) = \mathcal{L}_{\mathrm{SCE}} + \alpha\, \mathcal{L}_{\mathrm{DSC}} + \beta\, \mathcal{L}_{\mathrm{CM}}, \quad \alpha, \beta > 0. \tag{8}$$

We set the hyperparameters to $m_1 = 3$, $\alpha = 1.0$, $m_2 = 0.5$, and $\beta = 0.25$. When the model's accuracy on the backdoor class approaches random guessing, or a pre-

Algorithm 1 Logit Margin Repulsion (LMR)

Input: Backdoored model f_θ; clean set $\mathcal{D}_d$; margins $m_1, m_2 > 0$; weights $\alpha, \beta > 0$; $t = 0$; Phase-1 step budget T_1 (or early-switch rule); prune ratio p; Phase-2 steps T_2.
Output: Pruned model f_θ.

Save head weights $W^{(0)} \in \mathbb{R}^{K \times D}$.
Backdoor-class estimation. Estimate c on a small clean batch by *unlearning + log-posterior mean*; set $c \leftarrow \hat{y}_t$.
Phase-1: Logit-Geometric Purification
repeat
  Sample $(x, y) \sim \mathcal{D}_d$; logits $z \leftarrow f_\theta(x)$.
  Selective CE: $\mathcal{L}_{\mathrm{SCE}} = \mathbf{1}\{y \neq c\} \cdot \mathrm{CE}(z, y)$.
  Directional suppression (class c): $\mathcal{L}_{\mathrm{DSC}} = \big(z_c - \max_{j \neq c} z_j + m_1\big)_+ \cdot \mathbf{1}\{y \neq c\}$.
  Conditional margin (stability): $\mathcal{L}_{\mathrm{CM}} = \big(\max_{j \neq y} z_j - z_y + m_2\big)_+$.
  Total: $\mathcal{L}_{\mathrm{P1}} = \mathcal{L}_{\mathrm{SCE}} + \alpha\, \mathcal{L}_{\mathrm{DSC}} + \beta\, \mathcal{L}_{\mathrm{CM}}$.
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{P1}}$; $t \leftarrow t + 1$.
until $t \geq T_1$ **or** early-switch is satisfied (e.g., class-c clean accuracy $\approx$ random)
Save head weights at switch: $W^{(1)}$.
Phase-2: Delta-Prune + Light Recovery
Delta scores (target row only): $s_j = \big|W^{(1)}_{c,j} - W^{(0)}_{c,j}\big|$, $j = 1, \ldots, D$.
Column pruning: select $J = \text{Top-}k(\{s_j\})$ with $k = \lfloor pD \rfloor$; set $W_{:,J} \leftarrow 0$ and freeze their gradients.
for $u = 1$ **to** T_2 **do**
  Train with CE on $\mathcal{D}_d$; keep pruned columns frozen.
return pruned f_θ

defined number of epochs is reached, it proceeds to the next phase.

Phase 2: Delta-Prune & Light Recovery: We perform screening and pruning of the input channels of the linear classifier (head): using the ℓ_1 change in the backdoor-class weights of each channel before and after Phase-1 as a suspiciousness score, and prune the top k channels with the largest changes. This operation directly severs potential "feature $\rightarrow$ backdoor class" shortcut pathways, substantially reducing backdoor rebound during Phase 2 fine-tuning while keeping the decision boundaries of non-backdoor classes as stable as possible. Let the linear classifier (head) weights be $W \in \mathbb{R}^{K \times D}$. We store $W^{(0)}$ at the beginning of Phase 1 and $W^{(1)}$ right before switching to Phase 2. Let c denote the estimated backdoor class index. For each input column (feature) $j = 1, \ldots, D$, define the score:

$$s_j = \big|W^{(1)}_{c,j} - W^{(0)}_{c,j}\big|, \qquad j = 1, \ldots, D. \tag{9}$$

Subsequently, a small amount of data is used with standard cross-entropy loss to restore the model's accuracy on the backdoor class over clean data. See Algorithm 1 for the complete implementation of LMR.

4. Experiment

4.1. Experimental Setup

Datasets and Model Architectures: We conduct evaluations on CIFAR-10 [24], Tiny-ImageNet [25], and ImageNet [6], with resolutions of 32×32, 64×64, and 224×224, respectively. To verify the transferability of the proposed method, we conduct experiments[1] on ResNet [15], VGG [45], MobileNet-V2 [19, 44], and ViT [9].

Backdoor Attacks and Settings: Our experiments cover 12 attacks, including nine traditional backdoor attacks: BadNets [14], Blend [4], CL [50], SIG [29], Trojan [36], WaNet [40], LIRA [7], DFST [5], and Dynamic [41], as well as three conditional backdoor attacks (i.e., Quantized, Quantized-Distilled, and Pruned) [48]. For CIFAR-10 and Tiny-ImageNet, we inject backdoors into high-accuracy clean models obtained through training (for ViT, we directly adopt the official ViT-Base pretrained weights). The backdoor injection settings follow the original papers, and the trigger is fixed at the bottom-right corner by default, with sizes of 3×3 for CIFAR-10, 6×6 for Tiny-ImageNet, and 32×32 for ImageNet. The default poisoning rate is set to 10%. For QCB, we implement two attack variants: (i) a conventional QCB attack, which directly constructs a quantized backdoored model by exploiting rounding errors introduced during quantization; and (ii) a distillation-based QCB attack, which replaces hard labels with soft labels generated by a clean teacher model during training. For PCB attacks, we set the pruning ratio range to $[0.3, 0.9]$, within which the backdoor is activated (the default pruning ratio is set to 0.5 for metric evaluation in the experiments).

Backdoor Defenses and Settings: We compare LMR with eight backdoor defense methods, including FP [34], NAD [32], IAP [59], I-BAU [60], RNP [33], MNP [28], as well as two defense methods specifically designed for quantization-conditioned backdoors (QCB), namely EFRAP [27] and LACPDA [26]. The defense set is constructed from a randomly sampled 1% subset of the test dataset. For traditional backdoor attacks, all defense methods are applied to the original model, and the corresponding metrics are also computed on the original model. For QCB and PCB in conditional backdoor attacks, all defense methods are likewise conducted on the original model, while the corresponding metrics are computed on the quantized model (8-bit) and the pruned model (50% pruning), respectively.

[1] Our code is available on https://github.com/Trusted-LLM/LMR.

Table 2. Comparison with state-of-the-art defenses on CIFAR-10 with **1%** benign data on ResNet-18 . BadNets, Trojan, Blend, CL, SIG, WaNet, DFST, Dynamic, and LIRA are **TBA**. '⋆' denotes **CBA**: **QCB, QCB-D** and **PCB**. **Bold** indicates the best (highest ACC or lowest ASR), underline indicates the second best; methods marked with '†' are specialized for QCB and not counted in the main comparison.

Attack	No Defense		FP		NAD		IAP		I-BAU		RNP		MNP		EFRAP†		LACPDA†		LMR (Ours)	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Traditional backdoor attacks and defense performance																				
BadNets	96.36	96.00	95.67	4.04	**95.98**	1.44	93.80	9.31	93.91	4.37	92.54	0.76	94.10	0.52	96.32	96.14	95.87	83.38	95.85	**0.44**
Trojan	96.67	100.00	90.76	3.43	92.94	9.83	91.41	3.54	91.39	1.31	93.53	3.04	92.42	2.01	96.64	100.00	96.24	95.96	**94.45**	**0.93**
Blend	96.57	100.00	95.45	9.39	95.48	3.50	86.83	2.53	86.88	17.23	94.02	1.16	95.39	1.22	96.54	100.00	96.06	100.00	**95.59**	**0.28**
CL	94.91	93.76	87.77	8.17	**94.79**	32.93	85.85	2.17	88.90	31.23	93.23	3.47	91.14	0.61	95.06	95.31	94.70	63.60	94.65	**0.39**
SIG	96.39	99.73	81.55	8.21	89.66	11.19	91.56	11.23	86.53	38.66	93.18	0.59	94.13	0.50	96.35	99.77	95.77	92.68	**95.81**	**0.17**
WaNet	93.10	96.91	85.70	1.42	95.46	2.71	95.45	3.40	91.76	0.96	91.35	0.61	93.67	0.66	92.84	28.54	93.59	15.08	**95.50**	**0.44**
DFST	96.56	100.00	80.48	100.00	66.27	91.73	86.70	100.00	89.57	24.16	94.00	99.86	93.42	20.24	96.49	100.00	96.10	100.00	**94.57**	**0.70**
Dynamic	96.56	93.76	94.00	69.58	91.23	21.60	94.53	12.56	91.93	25.01	94.21	11.64	93.08	4.23	96.49	93.59	96.11	95.34	**94.60**	**1.12**
LIRA	92.87	91.06	91.92	17.53	86.13	33.44	91.90	4.73	90.54	0.56	90.91	0.87	91.56	0.76	92.84	91.31	91.86	86.70	**94.26**	**0.28**
Avg.(TBA)	95.55	96.80	89.26	24.64	89.77	23.15	90.89	16.61	90.16	15.94	93.00	13.56	93.21	3.42	95.51	89.41	95.14	81.42	**95.03**	**0.53**
Conditional backdoor attacks and defense performance																				
QCB⋆	91.62	99.73	90.98	84.00	91.22	1.12	82.14	99.43	89.81	10.24	90.64	0.54	86.26	0.70	91.46	0.60	**91.74**	1.07	91.15	**0.47**
QCB-D⋆	93.74	99.58	92.30	8.52	92.41	19.70	84.29	99.12	89.75	10.33	88.59	80.49	87.40	1.01	**93.78**	0.61	93.49	1.89	91.86	**0.14**
PCB⋆	81.74	99.99	83.59	83.01	78.22	27.34	80.69	6.84	**87.08**	76.17	83.26	7.56	81.52	4.71	90.17	99.38	90.15	99.27	84.08	**1.54**
Avg.(CBA)	89.03	99.77	88.96	58.51	87.28	16.05	82.37	68.46	88.88	32.25	87.50	29.53	85.06	2.14	91.80	33.53	91.79	34.08	**89.03**	**0.72**

Table 3. Comparison with the state-of-the-art defenses on ImageNet dataset with 1% benign data on ResNet-34 (%).

Attack	Backdoored		FP		NAD		IAP		I-BAU		RNP		MNP		LMR (Ours)	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	83.42	99.00	75.75	5.64	79.67	2.91	75.67	7.09	79.75	0.73	80.33	1.18	80.75	**0.64**	**81.75**	**0.64**
Trojan	83.42	100.00	75.25	9.91	75.83	1.27	75.42	6.73	80.50	1.09	79.83	0.73	79.67	0.82	**82.33**	**0.18**
Blend	81.75	100.00	73.58	21.45	77.83	1.55	75.75	2.36	77.08	90.36	76.33	2.82	81.08	1.55	**81.75**	**0.91**
CL	82.92	66.27	72.33	14.45	81.08	5.64	79.00	4.45	81.33	48.91	78.83	1.73	77.42	**1.27**	**82.50**	1.55
SIG	81.67	96.82	80.17	10.27	81.75	13.55	76.67	4.82	79.50	4.55	81.33	2.09	80.67	0.64	**82.17**	**0.55**
WaNet	83.25	98.45	79.58	9.36	78.67	15.91	81.50	20.45	79.92	18.27	78.67	0.45	79.58	1.00	**82.92**	**0.36**
LIRA	81.08	98.73	76.17	11.00	79.58	0.91	79.50	2.36	78.25	51.09	81.67	0.73	**82.42**	0.73	**82.42**	**0.55**
Average	82.50	94.18	76.12	11.73	79.20	5.96	77.64	6.89	79.48	30.71	79.57	1.39	80.23	0.95	**82.26**	**0.68**

Evaluation Metrics: Clean accuracy (ACC) and attack success rate (ASR) [4] are the two most commonly used metrics for evaluating backdoor defenses. In addition, we also introduce trigger accuracy (TA). An ideal purified model should achieve high ACC and TA while maintaining a low ASR.

4.2. Experimental Results

Traditional Backdoor Attack: We evaluate the defense performance of LMR against both traditional and conditional backdoor attacks on CIFAR-10 and ImageNet using ResNet-18 / 34. Here, "No Defense" denotes the original backdoored model, while FP, NAD, IAP, I-BAU, RNP, MNP, EFRAP, and LACPDA serve as defense baselines. As shown in Tabs. 2 and 3, existing defense methods exhibit clear limitations in certain scenarios: FP / NAD perform poorly under content-aware attacks; RNP / MNP have limited effectiveness against conditional backdoor attacks; moreover, EFRAP / LACPDA are primarily designed for *QCB* and thus are ineffective for defending against traditional backdoor attacks. With only 1% clean data, LMR achieves competitive results across different datasets and attack settings compared with various baselines. In the TBA scenario, LMR effectively removes backdoors with only a small accuracy cost: on CIFAR-10, the average ASR decreases from 96.80% to 0.53%, with only a 0.5% drop in ACC; on ImageNet, the average ASR decreases from 94.2% to 0.68%, with only a 0.25% drop in ACC.

In addition, as shown in Fig. 3, in the original backdoored model, samples from different classes form rela-

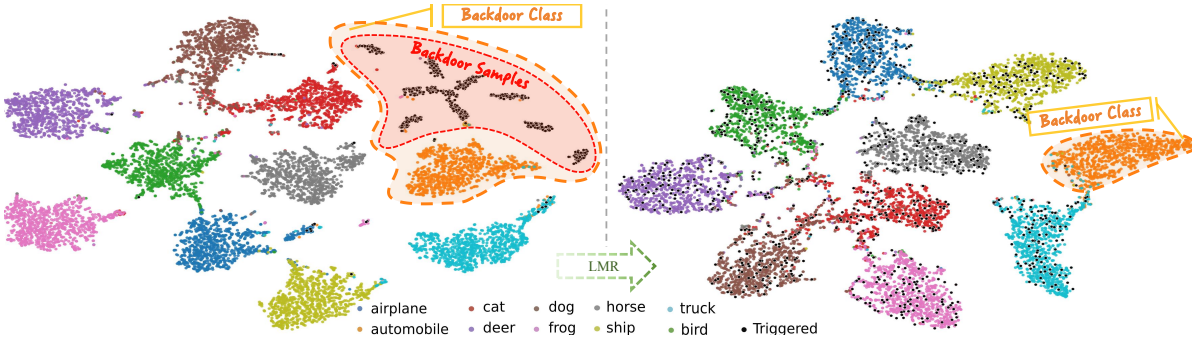

Figure 3. The t-SNE plots of the feature distribution of samples from different classes, in the backdoored model, trigger samples cluster toward the backdoor class, whereas after applying LMR, they move back to the neighborhood of their source classes.

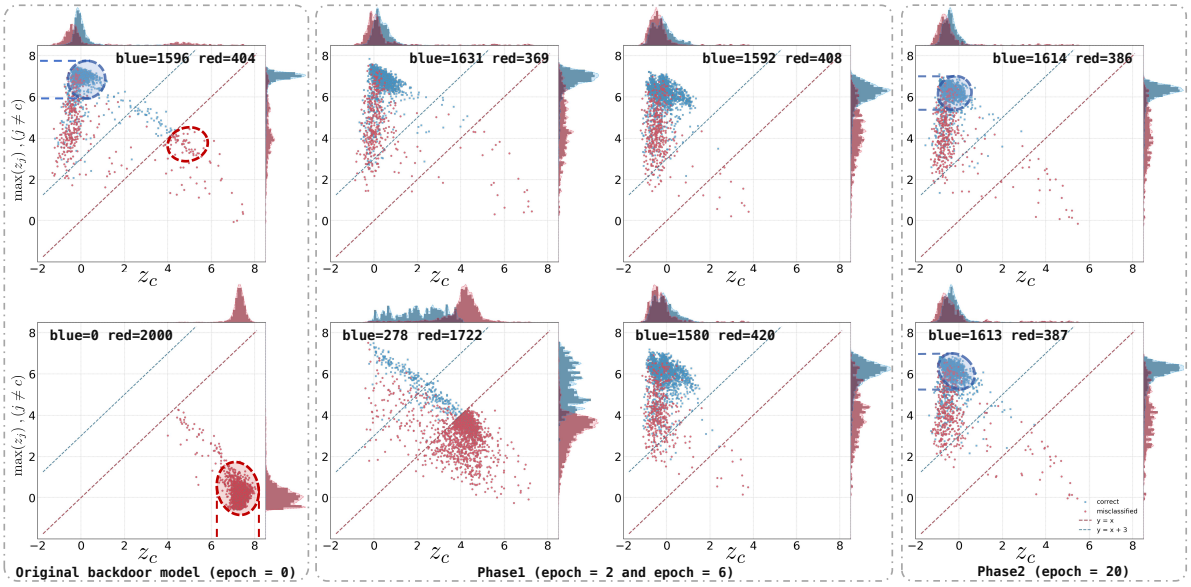

Figure 4. Visualization of the backdoor purification process in the $(z_c, \hat{z})$. The x-axis is the backdoor-class logit z_c; the y-axis is the maximum non-backdoor logit $\hat{z} = \max_{j \neq c} z_j$. The black dashed line $\hat{z} = z_c$ is the decision boundary: points above it are in the safe region and points below it are in the risk region. The purple dashed line $\hat{z} = z_c + m_1$ indicates the desired safety margin (we set $m_1 = 3$ in our experiments).

tively separated clusters in the feature space, while trigger-bearing samples are distributed near the backdoor-class manifold and drift across multiple clusters. After LMR purification, benign samples still maintain clear and compact intra-class structures, and trigger-bearing samples no longer collapse toward the backdoor class but instead return to the neighborhood of their source classes.

In Fig. 4, each subfigure shows the distribution of 2,000 non-backdoor-class samples, where the red highlighted points denote all misclassified samples among the 10,000 test samples. The first row corresponds to the distribution of clean samples, and the second row corresponds to the distribution of triggered samples. In the original backdoored model (epoch 0), the trigger significantly increases the logit corresponding to the backdoor class, causing all triggered samples to be predicted as the target class, with the red points densely distributed below the decision boundary. Meanwhile, a small number of clean samples are distributed near the boundary. After Phase-1 (epochs 2 and 6), the distribution of triggered samples shifts upward as a whole, and most samples are correctly classified, gradually aligning with the predefined safety margin. As the defense proceeds, regardless of whether the trigger is present, the vast majority of sample points become stably distributed above the predefined safety margin; at the same time, the distributions of triggered and clean samples gradually become more concentrated, and TA progressively approaches ACC. On the other hand, the overall distribution of clean samples remains close to that of the original model, indicating that the model's discriminative ability for non-backdoor classes is well preserved.

Conditional Backdoor Attack: As shown in Tab. 2, LMR also demonstrates outstanding defense performance

Table 4. Defense efficacy of **LMR** against 8-bit *QCB* attacks on **CIFAR-10** (*No Defense = ND*).

Setting	Metric	ResNet-18				VGG-16				MobileNetV2			
		ND	EFRAP	LACPDA	**LMR**	ND	EFRAP	LACPDA	**LMR**	ND	EFRAP	LACPDA	**LMR**
No-Distilling	ACC↑	91.62	91.46	91.74	91.15	89.50	84.34	89.58	88.53	91.58	91.31	91.87	91.47
	ASR↓	99.81	0.60	1.10	**0.47**	100.00	**0.22**	0.57	0.54	92.52	2.70	0.47	**0.01**
Distilling	ACC↑	93.83	93.78	93.43	91.86	91.65	92.35	92.39	91.68	92.05	92.35	91.73	91.47
	ASR↓	99.53	0.61	1.89	**0.14**	99.94	0.70	0.56	**0.49**	99.39	0.93	**0.57**	**0.57**

Table 5. Ablation study on each component, "CE" denotes using only the standard cross-entropy loss.

Model - Dataset	Metric	No Defense	CE	SCE + DSC $m_1 = \{1, 2, 5, 10\}$				SCE + CM $m_2 = 0.5$	SCE + DSC + CM $m_1 = 10,\ m_2 = 0.5$
ResNet18 - CIFAR10	ACC↑	95.84	96.43	96.32	96.24	96.23	96.37	96.28	**96.31**
	ASR↓	98.92	92.70	69.57	42.77	35.97	4.84	9.83	**0.69**
ViT - ImageNet	ACC↑	94.33	90.75	91.58	90.33	90.00	92.58	90.33	***92.58***
	ASR↓	99.82	93.09	83.82	21.36	2.91	3.36	64.91	**0.91**

against conditional backdoor attacks. For QCB, the ASR decreases from 99.65% to 0.3%, with only approximately 1% degradation in accuracy. For the original PCB model, after 50% pruning, the ACC drops from 90.93% to 81.74%, while the ASR surges from 0.40% to 99.99%; however, after applying LMR, the ACC recovers to 84.08%, and the ASR is significantly reduced to 1.54%. Furthermore, as shown in Tab. 4, we further evaluate the defense capability of LMR against QCB on ResNet-18, VGG-16, and MobileNet-V2. The results show that LMR can consistently remove backdoors while maintaining high clean accuracy, demonstrating good generalization ability.

4.3. Ablation Studies

Impact of Defense Data: We evaluate the impact of defense data size on LMR on CIFAR-10 / ResNet-18, using BadNets (visible trigger) and LIRA (invisible trigger) as representative attack scenarios. As shown in Fig. 5, even with only **0.02%** defense data, LMR is able to reduce the ASR of common backdoor attacks to approximately $\approx 0.5\%$; as the defense set grows, LMR can further improve the model's ACC, making its output distribution closer to that of the original clean model.

Ablation Study on Loss Terms: To analyze the role of each loss term, we adopt a low learning rate ($1e{-}3$) and use 0.6% clean samples, ensuring that CE alone cannot effectively suppress the backdoor. Experiments are conducted on CIFAR-10 / ResNet-18 (BadNets). We compare three main loss configurations: CE only, SCE+DSC, and SCE+DSC+CM; except for the "CE only" setting, SCE is included by default in the other settings. As shown in Tab. 5, when using CE alone, the ASR decreases slowly. After introducing the margin-controlled DSC, the ASR is significantly reduced, and a larger margin m_1 leads to a lower

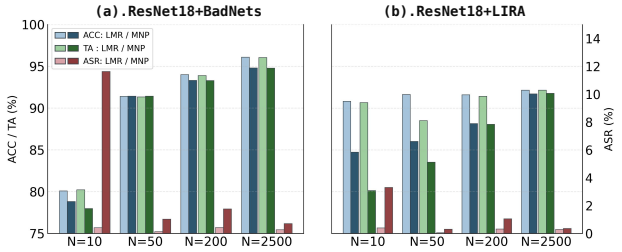

Figure 5. A comparison of **LMR** and **MNP** against *BadNets* and *LIRA* on `ResNet-18/CIFAR-10`, with the defense data size N varying from 0.02% (10) to 5% (2500) of the training set.

ASR. Further adding the CM term improves the stability of the model. In addition, we analyze the sensitivity of the hyperparameters α and β. The parameter α controls the suppression strength on the backdoor logit and requires only *coarse-grained selection*: when $\alpha \in [0.5, 3]$, ACC and ASR remain within $[83.7, 94.4]$ and $[0.11, 0.36]$, respectively. In contrast, LMR is *insensitive* to β: when $\beta \in [0.1, 1.0]$, ACC and ASR remain within $[92.8, 94.4]$ and $[0.27, 0.30]$, respectively.

5. Conclusion

We propose LMR, a direct and effective defense against backdoor attacks. Extensive experiments show that LMR requires only a small amount of clean data while achieving highly competitive defense performance against both traditional and conditional backdoors. Its core lies in imposing a margin constraint in the logit space (DSC), together with selective cross-entropy (SCE) and conditional margin (CM), to suppress backdoor behaviors. As a logit-based method, LMR exhibits strong practicality and scalability. Although more sophisticated attacks may emerge in the future, our results show that, under current backdoor threat scenarios, LMR is an efficient and general defense method.

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019. 1, 2

[2] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018. 2

[3] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, page 8, 2019. 2

[4] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2, 5, 6

[5] Siyuan Cheng, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. Deep feature space trojan attack of neural networks by controlled detoxification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1148–1156, 2021. 2, 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5

[7] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11966–11976, 2021. 5

[8] Peiran Dong, Haowei Li, and Song Guo. Durable quantization conditioned misalignment attack on large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[10] Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting llm quantization. *Advances in Neural Information Processing Systems*, 37:41709–41732, 2024. 2

[11] Kazuki Egashira, Robin Staab, Mark Vero, Jingxuan He, and Martin Vechev. Mind the gap: A practical attack on gguf quantization. *arXiv preprint arXiv:2505.23786*, 2025. 2

[12] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021. 1

[13] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4852–4861, 2019. 2

[14] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 2, 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Sanghyun Hong, Michael-Andrei Panaitescu-Liess, Yigitcan Kaya, and Tudor Dumitras. Qu-anti-zation: Exploiting quantization artifacts for achieving adversarial outcomes. *Advances in Neural Information Processing Systems*, 34:9303–9316, 2021. 1, 2, 3

[17] Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. *arXiv preprint arXiv:2405.09786*, 2024. 2, 4

[18] Sizai Hou, Songze Li, and Duanyi Yao. Dede: Detecting backdoor samples for ssl encoders via decoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20675–20684, 2025. 2

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[20] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020. 1

[21] Xiaowei Huang, Gaojie Jin, and Wenjie Ruan. *Machine learning safety*. Springer, 2023.

[22] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7): 175, 2024. 1

[23] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 301–310, 2020. 2

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5

[25] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[26] Boheng Li, Yishuo Cai, Jisong Cai, Yiming Li, Han Qiu, Run Wang, and Tianwei Zhang. Purifying quantization-conditioned backdoors via layer-wise activation correction with distribution approximation. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 5

[27] Boheng Li, Yishuo Cai, Haowei Li, Feng Xue, Zhifeng Li, and Yiming Li. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24523–24533, 2024. 2, 3, 5

[28] Nan Li, Haoyu Jiang, and Ping Yi. Magnitude-based neuron pruning for backdoor defens. *arXiv preprint arXiv:2405.17750*, 2024. 3, 5

[29] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5): 2088–2105, 2020. 5

[30] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021. 2

[31] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021. 2

[32] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021. 2, 5

[33] Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. In *International Conference on Machine Learning*, pages 19837–19854. PMLR, 2023. 1, 2, 3, 5

[34] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018. 2, 3, 5

[35] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *2017 IEEE international conference on computer design (ICCD)*, pages 45–48. IEEE, 2017. 2

[36] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018. 5

[37] Hua Ma, Huming Qiu, Yansong Gao, Zhi Zhang, Alsharif Abuadbba, Minhui Xue, Anmin Fu, Jiliang Zhang, Said F Al-Sarawi, and Derek Abbott. Quantization backdoors to deep learning commercial frameworks. *IEEE Transactions on Dependable and Secure Computing*, 21(3):1155–1172, 2023. 2

[38] Naren Manoj and Avrim Blum. Excess capacity and backdoor poisoning. *Advances in Neural Information Processing Systems*, 34:20373–20384, 2021. 1, 2

[39] Ronghui Mu, Wenjie Ruan, Leandro S Marcolino, and Qiang Ni. 3dverifier: efficient robustness verification for 3d point cloud models. *Machine Learning*, 113(4):1771–1798, 2024. 1

[40] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 5

[41] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020. 5

[42] Xudong Pan, Mi Zhang, Yifan Yan, and Min Yang. Understanding the threats of trojaned quantized neural network in model supply chains. In *Proceedings of the 37th Annual Computer Security Applications Conference*, pages 634–645, 2021. 2

[43] Neehar Peri, Neal Gupta, W Ronny Huang, Liam Fowl, Chen Zhu, Soheil Feizi, Tom Goldstein, and John P Dickerson. Deep k-nn defense against clean-label data poisoning attacks. In *European Conference on Computer Vision*, pages 55–70. Springer, 2020. 2

[44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[46] Siqi Sun, Procheta Sen, and Wenjie Ruan. Crowd: Certified robustness via weight distribution for smoothed classifiers against backdoor attack. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17056–17070, 2024. 1

[47] Guanhong Tao, Guangyu Shen, Yingqi Liu, Shengwei An, Qiuling Xu, Shiqing Ma, Pan Li, and Xiangyu Zhang. Better trigger inversion optimization in backdoor scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13368–13378, 2022. 2

[48] Yulong Tian, Fnu Suya, Fengyuan Xu, and David Evans. Stealthy backdoors as compression artifacts. *IEEE Transactions on Information Forensics and Security*, 17:1372–1387, 2022. 1, 2, 5

[49] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018. 2

[50] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018. 5

[51] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019. 2

[52] Fu Wang, Peipei Xu, Wenjie Ruan, and Xiaowei Huang. Towards verifying the geometric robustness of large-scale neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 15197–15205, 2023. 1

[53] Hang Wang, Zhen Xiang, David J Miller, and George Kesidis. Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1994–2012. IEEE, 2024. 2

[54] Tong Wang, Yuan Yao, Feng Xu, Miao Xu, Shengwei An, and Ting Wang. Inspecting prediction confidence for detecting black-box backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 274–282, 2024. 2, 4

[55] Zhenting Wang, Kai Mei, Hailun Ding, Juan Zhai, and Shiqing Ma. Rethinking the reverse-engineering of trojan triggers. *Advances in Neural Information Processing Systems*, 35:9738–9753, 2022. 2

[56] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 2

[57] Han Wu, Syed Yunas, Sareh Rowlands, Wenjie Ruan, and Johan Wahlström. Adversarial driving: Attacking end-to-end autonomous driving. In *2023 IEEE intelligent vehicles symposium (IV)*, pages 1–7. IEEE, 2023. 1

[58] Peipei Xu, Wenjie Ruan, and Xiaowei Huang. Quantifying safety risks of deep neural networks. *Complex & Intelligent Systems*, 9(4):3801–3818, 2023. 1

[59] Mingfu Xue, Yinghao Wu, Zhiyu Wu, Yushu Zhang, Jian Wang, and Weiqiang Liu. Detecting backdoor in deep neural networks via intentional adversarial perturbations. *Information Sciences*, 634:564–577, 2023. 5

[60] Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021. 3, 5

[61] Chi Zhang, Wenjie Ruan, and Peipei Xu. Reachability analysis of neural network control systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15287–15295, 2023. 1

[62] Zhendong Zhao, Xiaojun Chen, Yuexin Xuan, Ye Dong, Dakui Wang, and Kaitai Liang. Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15213–15222, 2022. 2

[63] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision*, pages 175–191. Springer, 2022. 2

[64] Feng Zhu, Ruihao Gong, Fengwei Yu, Xianglong Liu, Yanfei Wang, Zhelong Li, Xiuqi Yang, and Junjie Yan. Towards unified int8 training for convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1979, 2020. 2

[65] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4466–4477, 2023. 2