1.

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Huy Phan, *Jinqi Xiao, Yang Sui, Tianfang Zhang , Zijie Tang,*
Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan

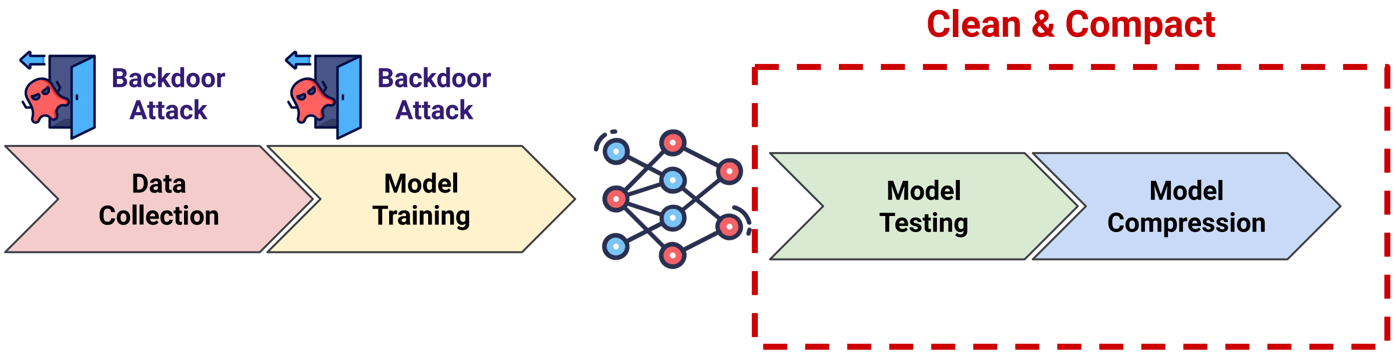
[ECCV-24] European Conference on Computer Vision 2024

2.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Threat Model



- **Input**: Given a trained model that potentially has backdoors.
- **Task:** Remove the potential backdoors **and** simultaneously compress the model size for resource-constrained device.

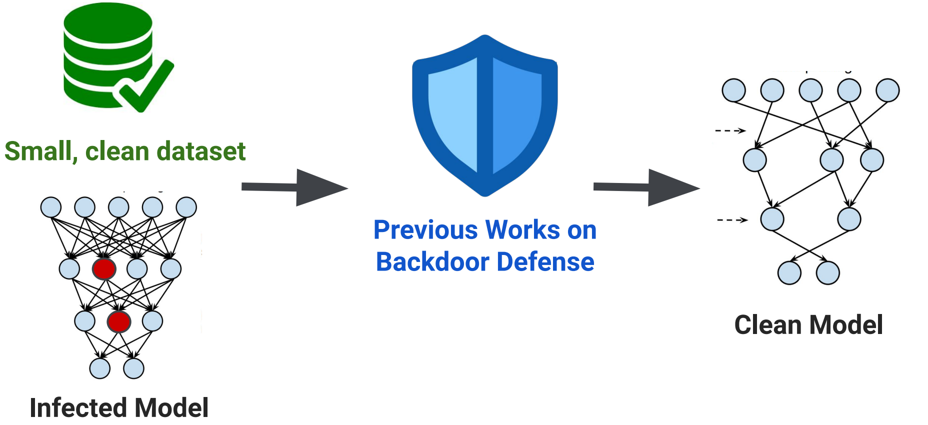
3.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Previous Works on Backdoor Defense

- **Given** a model infected with backdoors, try to identify the infected parts of model (neurons, channels) then prune them.
- **Requires** 1%-5% clean data to identify infected parts of the model, and fine-tune the model after pruning.



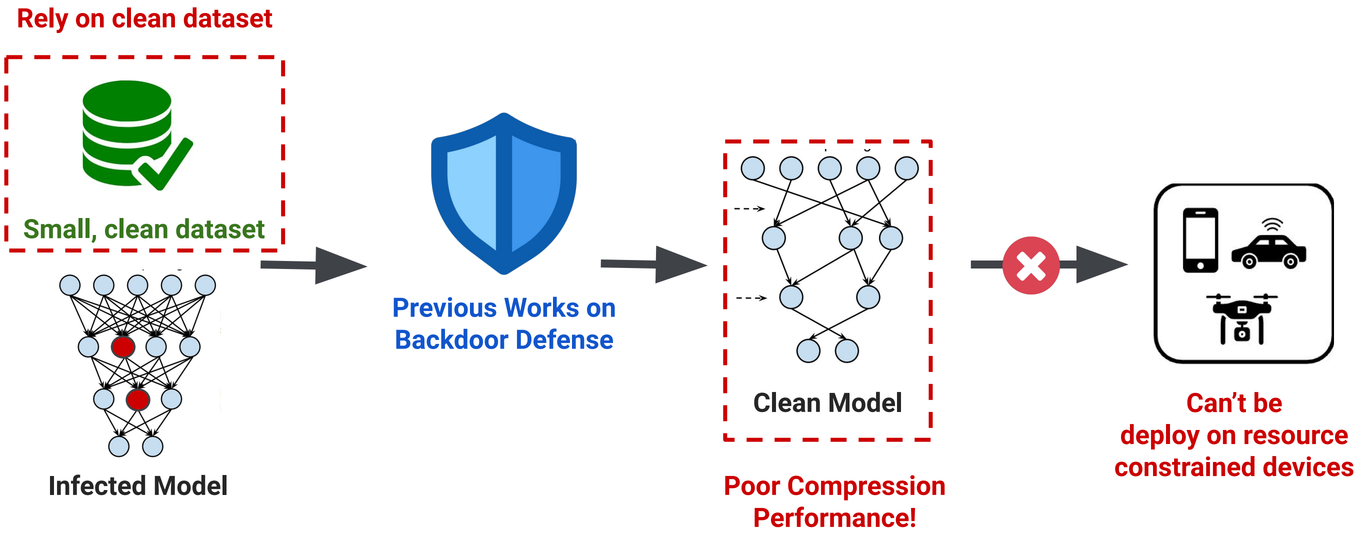
4.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Previous Works on Backdoor Defense

- **Given** a model infected with backdoors, try to identify the infected parts of model (neurons, channels) then prune it.
- **Requires** 1%-5% clean data to identify infected parts of the model, and fine-tune the model after pruning.



5.

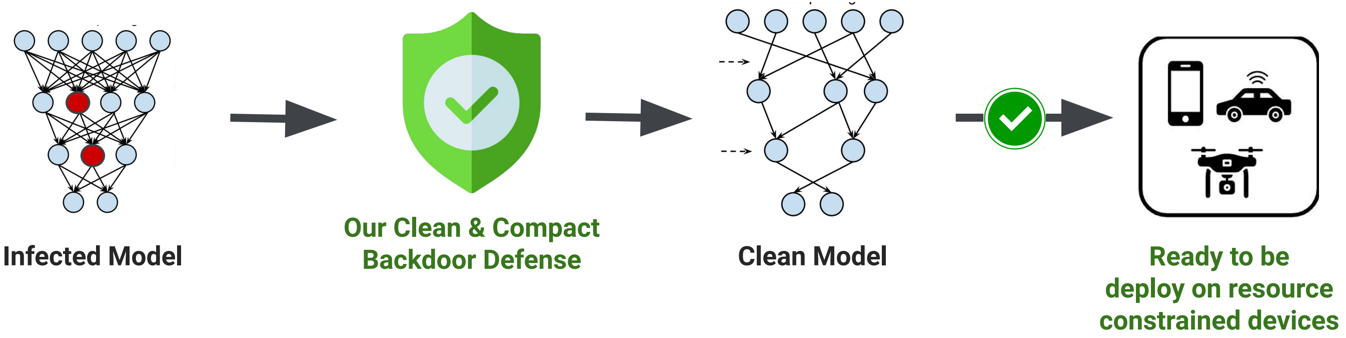
[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Our motivation

Develop a backdoor defense that can simultaneously:

- Effectively remove backdoors from infected model.
- Achieve high compression performance.
- Do not rely on any data at all.

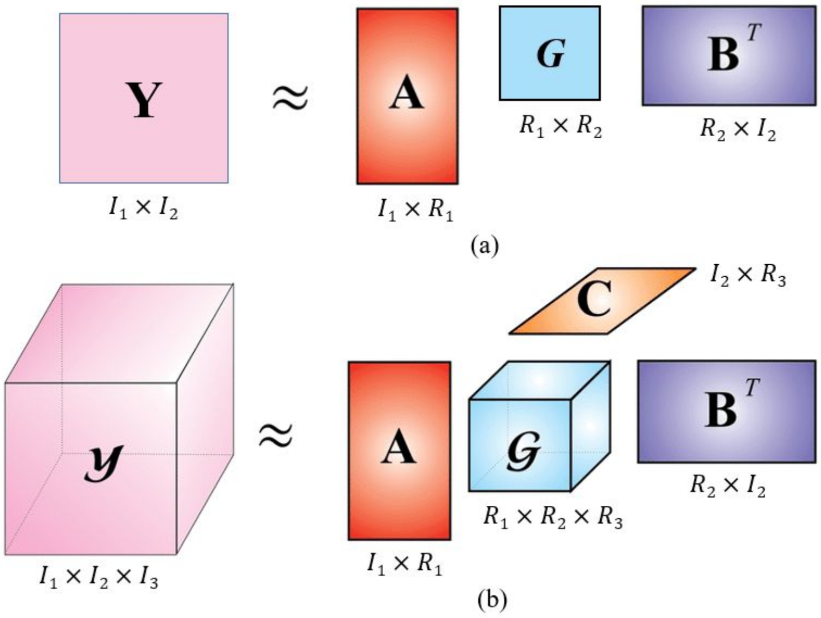


6

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Low Rank Decomposition



Matrix decomposition vs. tensor decomposition:
(a) low-rank matrix decomposition (truncated SVD);
(b) low-rank tensor decomposition (Tucker decomposition).

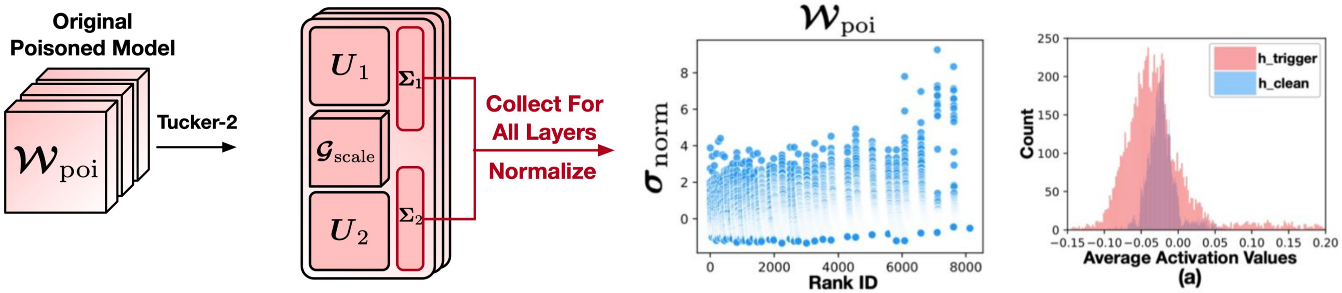
7.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Observation

- **Key Idea: Explore Model Backdoor Sensitivity From Singular Values**
- Decompose all weight tensor using Tucker-2, collect the singular values of all layers.
- Plot the normalized singular values, together with activation values of ***backdoor examples***, and ***clean examples***



8.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Observation

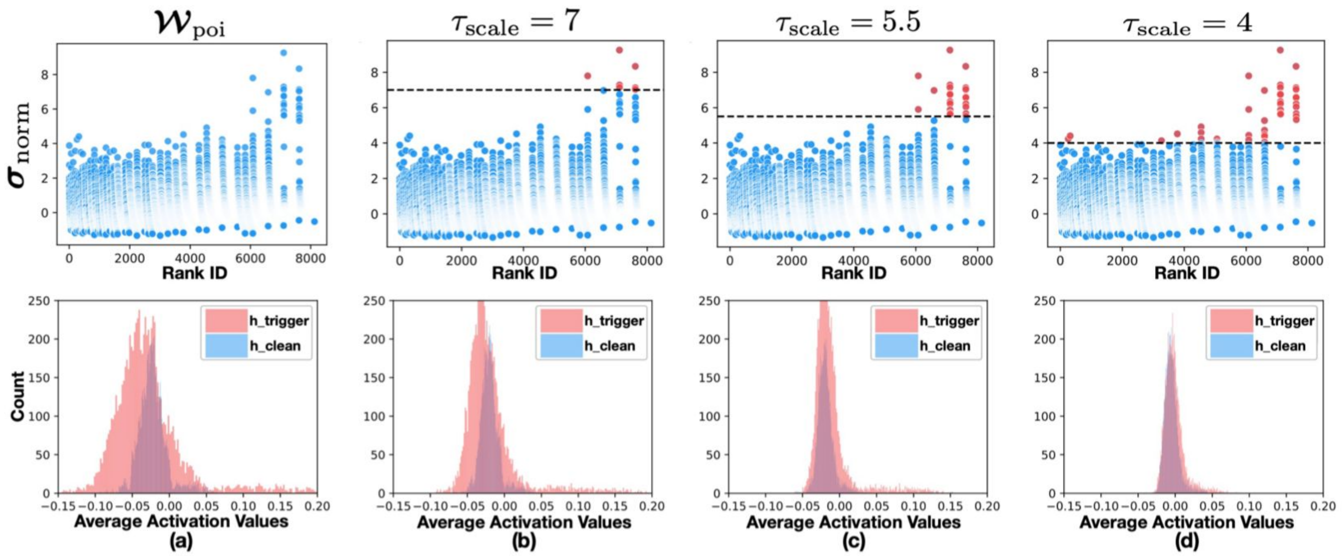


Fig. 1: (1st Row) Decreasing τ_{scale} makes more high-valued normalized singular values being scaled down. (2nd Row) As τ_{scale} decreases, $\boldsymbol{h}_{trigger}$ shrinks to approach $\boldsymbol{h}_{clean}$. The model architecture is ResNet-18 on CIFAR-10 and the backdoor attack is WaNet.

9.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Removing backdoors, and reducing model size

1) Apply Tucker-2 low rank decomposition:

$$\mathcal{W} = \mathcal{G} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2,$$

2) Scale the rank components in mode-1 matricization:

$$\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times K \times K} \xrightarrow{\text{unfold}} \boldsymbol{G}_{(1)} \in \mathbb{R}^{r_1 \times (r_2 * K * K)},$$
$$\boldsymbol{G}_{(1)}^{\text{scale}} = \boldsymbol{G}_{(1)} \odot \min(\tau_{\text{scale}} * s_{\boldsymbol{\sigma}} / \boldsymbol{T}_1, 1),$$

3) Scale the rank components in mode-2 matricization:

$$\boldsymbol{G}_{(1)}^{\text{scale}} \in \mathbb{R}^{r_1 \times (r_2 * K * K)} \xrightarrow{\text{reshape}} \boldsymbol{G}_{(2)}^{\text{temp}} \in \mathbb{R}^{r_2 \times (r_1 * K * K)},$$
$$\boldsymbol{G}_{(2)}^{\text{scale}} = \boldsymbol{G}_{(2)}^{\text{temp}} \odot \min(\tau_{\text{scale}} * s_{\boldsymbol{\sigma}} / \boldsymbol{T}_2, 1),$$

4) Prune the ranks component to reduce the model size:

$$\boldsymbol{G}_{(2)}^{\text{scale}} \xrightarrow{\text{fold}} \mathcal{G}_{\text{scale}} \in \mathbb{R}^{r_1 \times r_2 \times K \times K}, \quad \text{and} \quad \mathcal{W}_{\text{constrain}} = \mathcal{G}_{\text{scale}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2. \quad (7)$$

$$\mathcal{W}_{\text{constrain}}^{\text{comp}} = \mathcal{G}_{\text{scale}}^{\text{comp}} \times_1 \boldsymbol{U}_1^{\text{comp}} \times_2 \boldsymbol{U}_2^{\text{comp}}, \text{where} \begin{cases} \mathcal{G}_{\text{scale}}^{\text{comp}} = \mathcal{G}_{\text{scale}}(1:R_1, 1:R_2), \\ \boldsymbol{U}_1^{\text{comp}} = \boldsymbol{U}_1(1:R_1), \\ \boldsymbol{U}_2^{\text{comp}} = \boldsymbol{U}_2(1:R_2). \end{cases}$$

10.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Boosting Performance via Synthetic Data-Aided Fine-Tuning

- Considering the unavailability of training dataset in the realistic data-free setting, we propose to generate ***synthetic*** data for efficient fine-tuning.
- Synthetic data can be generated by ZeroQ algorithm by iteratively optimizing randomly generated data to match model's batch norms stats

$$\min_{\boldsymbol{x}_s} \sum_{j=1}^{L} ||\tilde{\mu}_j^s - \mu_j||_2^2 + ||\tilde{\sigma}_j^s - \sigma_j||_2^2 + \mathcal{L}(F_{\{\mathcal{W}^{\text{comp}}_{\text{constrain}}\}}(x_s), \boldsymbol{y}),$$

- When performance ***untarget*** adversarial attacks on these synthetic data, most adv. examples fall into the backdoor class.
- Hence these adversarial synthetic data can serve as a proxy for real backdoor data for fine-tuning

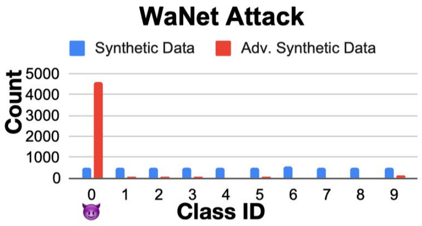


Fig. 4: Generated from syn. data with added adv. noise, most are labeled to target (class-0), implying they can serve as surrogates for real poisoned data.

11.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Overall process of Clean & Compact

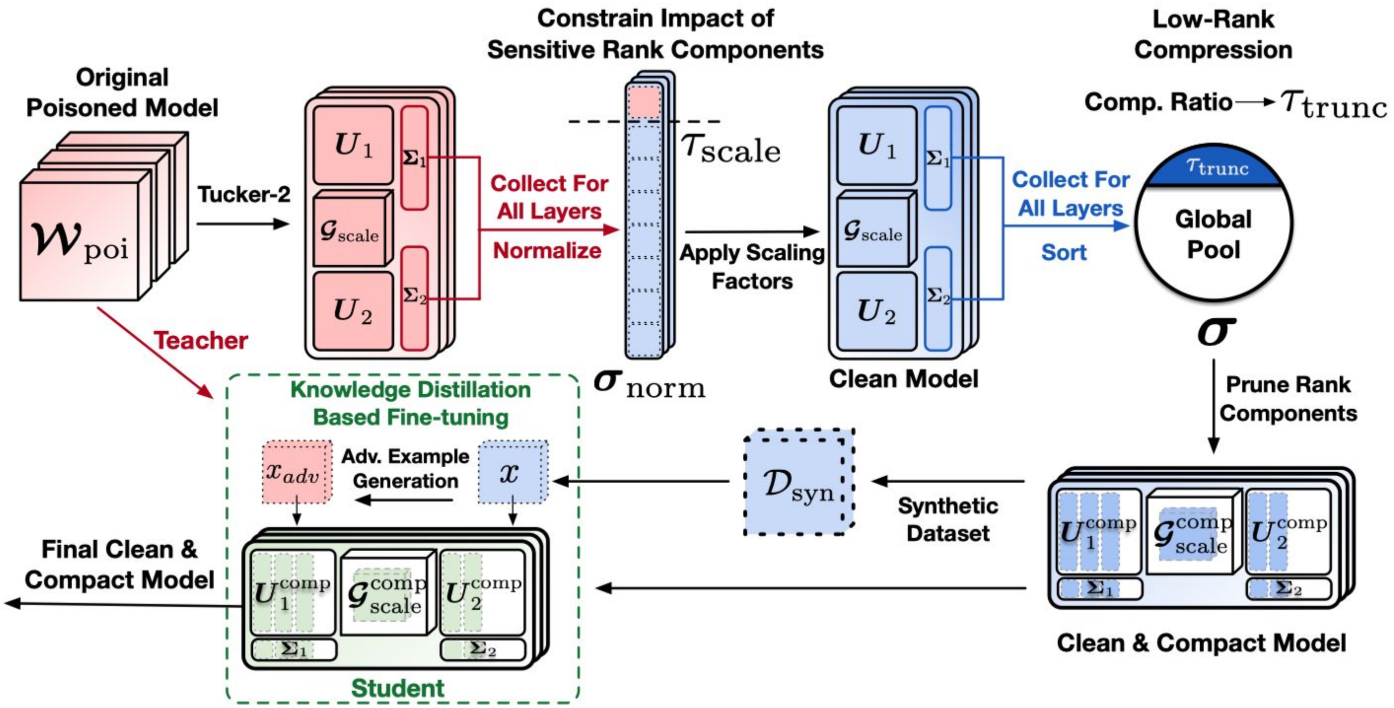


Fig. 2: The overall process of obtaining a data-free, clean and compact DNN.

12.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Table 2: Performance for jointly purifying and compressing ResNet-18 on CIFAR-10. ACC of ANP/CLP drops to 10% with 2× compression. C&C maintains high ACC from 2× to 4× compression, showing superior performance at higher ratios, being data-free. Inference time is measured on a NVIDIA RTX 3090 GPU.

	Defense Methods - Compression Ratio											
	No Defense		**ANP 2×**		**CLP 2×**		**C&C 2×**		**C&C 3×**		**C&C 4×**	
Attacks↓	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNet	94.13	97.96	10.00	0.00	10.00	0.00	92.16	2.88	91.25	1.30	90.77	0.71
Blended	93.45	99.67	10.00	0.00	11.30	0.00	91.01	4.02	90.48	2.49	89.13	2.54
InputAware	94.33	99.60	23.05	25.78	10.00	0.00	92.93	0.90	92.84	0.70	92.70	0.60
WaNet	93.71	99.32	10.00	0.00	10.00	0.00	92.38	1.41	92.72	2.40	92.28	1.10
BadNet A2A	93.70	91.12	12.38	10.21	10.00	10.00	92.42	3.87	91.85	3.66	91.27	3.36
Blended A2A	93.59	92.59	10.00	10.00	10.00	10.00	90.78	5.62	90.00	4.81	90.14	4.52
InputAware A2A	94.01	91.79	10.00	10.00	13.68	11.72	93.46	1.80	93.06	2.30	91.19	2.49
WaNet A2A	93.74	92.18	10.00	10.00	10.00	10.00	93.16	2.03	92.82	1.81	91.83	1.84
Data Req.	N/A		1% clean		Data-free		Data-free		Data-free		Data-free	
Comp. Type	N/A		Unstructured		Channel		Low-rank		Low-rank		Low-rank	
Parameters	11.17M		5.58M		5.58M		5.58M		3.72M		2.78M	
Inference Time	0.201ms		0.201ms		0.150ms		**0.143ms**		**0.125ms**		**0.110ms**	
Speed Up	N/A		None		1.34×		**1.41×**		**1.61×**		**1.83×**	

CLP [43]. Here except CLP adopting data-free defense strategy, NAD, ANP and I-BAU are set to have access to the same 1% clean training data.

13.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Table 4: Backdoor defense performance across different datasets using ResNet-18.

		No Defense		CLP		C&C (Ours)	
Datasets	**Attacks**	**ACC**	**ASR**	**ACC**	**ASR**	**ACC**	**ASR**
GTSRB	BadNet	97.17	97.20	98.70	8.52	97.70	2.96
	BadNet A2A	98.97	95.40	97.65	0.48	96.32	5.76
	InputAware	98.99	98.81	98.85	7.72	98.94	0.00
	InputAware A2A	98.45	96.97	95.87	15.61	98.59	0.14
	Average	98.40	97.10	97.77	8.08	**97.89**	**2.22**
CIFAR-100	BadNet	74.35	96.71	44.78	0.81	70.27	1.83
	BadNet A2A	74.15	69.40	53.20	0.88	73.28	0.95
	InputAware	65.49	93.92	53.92	6.59	60.58	6.19
	InputAware A2A	66.19	57.13	53.57	0.87	64.12	5.22
	Average	70.05	79.29	51.37	**2.29**	**67.06**	3.55

14

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

	CIFAR-10						GTSRB					
	No Defense		CLP		C&C		No Defense		CLP		C&C	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNet Attack												
ResNet-34	90.13	97.94	83.61	0.58	89.34	0.94	97.84	98.20	97.70	7.61	97.95	0.48
VGG-19	89.68	95.83	83.25	1.38	89.15	3.08	97.42	94.91	96.67	5.62	97.55	0.35
MobileNet-V2	89.56	86.26	83.61	0.58	87.10	1.10	96.86	96.52	92.41	0.03	97.16	1.23
Average	89.79	93.34	83.49	**0.85**	**88.53**	1.71	97.37	96.54	95.59	4.42	**97.55**	**0.69**
InptutAware Attack												
ResNet-34	91.67	86.98	85.64	2.12	89.46	0.95	98.59	94.40	98.76	0.50	98.54	0.15
VGG-19	89.01	82.39	85.64	2.12	89.03	1.30	97.28	91.60	95.76	0.28	97.14	0.06
MobileNet-V2	89.45	82.38	80.53	2.93	88.93	1.42	97.64	93.78	95.86	1.29	96.89	1.58
Average	90.04	83.92	83.94	2.39	**89.14**	**1.22**	97.84	93.26	96.79	0.69	**97.52**	**0.60**

15.

[ECCV-24]

Clean & Compact: Efficient Data-Free Backdoor Defense with Model Compactness

Conclusion

- We develop a backdoor defense that can effectively remove backdoors, achieve high compression performance without using any data.
- Overall, the Clean & Compact (C&C) method addresses critical gaps in backdoor defense, paving the way for more secure and efficient deployment of DNNs across various applications.