IBD-PSC: Input-level Backdoor Detection via Parameter-oriented Scaling Consistency

Linshan Hou, Ruili Feng, Zhongyun Hua*, Weo Luo, Leo Yu Zhang, Yiming Li*

huazhongyun@hit.edu.cn; liyiming.tech@gmail.com

Get paper here

Get code here

Introduction

➢ We proposes a simple yet effective input-level backdoor detection (dubbed **IBD-PSC**) as **a 'firewall' to filter out malicious testing images**.

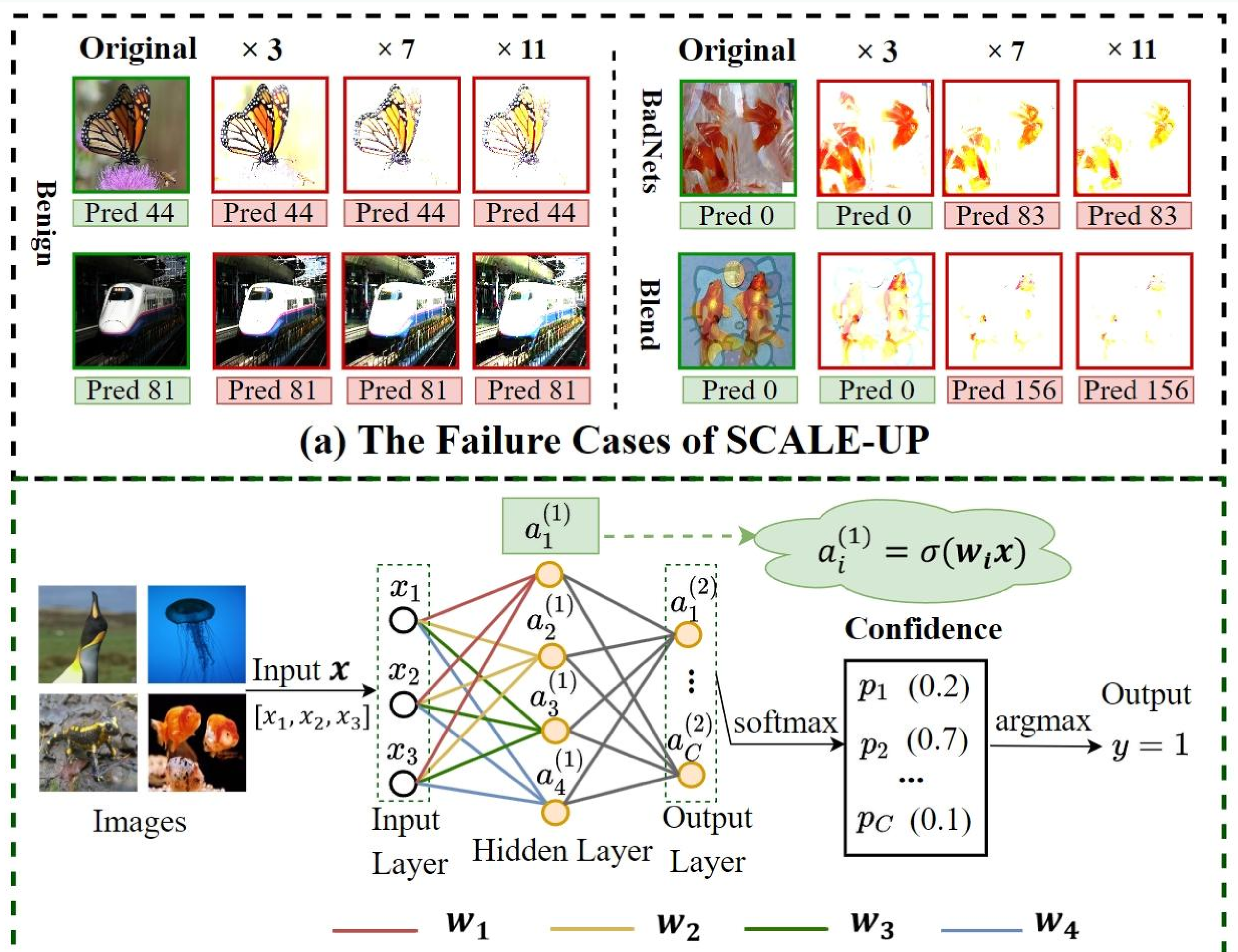
Motivation

➢ The most advanced IBD method, SCALE-UP, encounters intrinsic limitations (as shown in Fig. 1(a)) due to the restriction of pixel values (i.e., bounded in [0, 255]).

➢ The predictions are from the co-effects of pixel and parameter values, as shown in Fig. 1(b).

➢ Parameter values are not bounded.

Shall the model's parameters expose backdoors with more grace than the humble pixel's tale?



(a) The Failure Cases of SCALE-UP

(b) The Co-effects of Pixel and Parameter Values

Main Contributions

➢ We disclose the parameter-oriented scaling consistency (PSC) phenomenon, where the prediction confidences of poisoned samples are more consistent than benign ones when scaling up BN parameters.

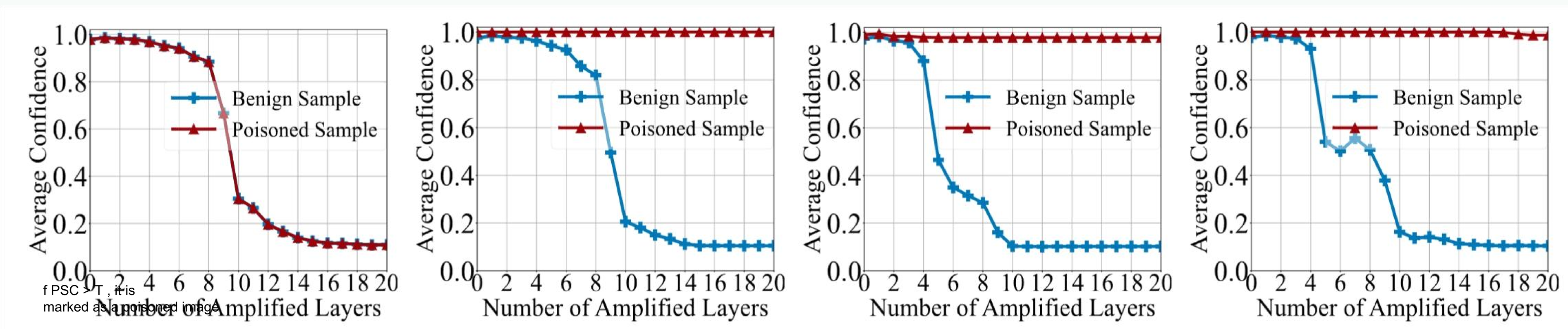
➢ We provide theoretical insights to elucidate the PSC phenomenon.

➢ We design a simple yet effective method (i.e., IBD-PSC) to filter out poisoned testing images based on our findings.

➢ Extensive experiments on benchmark datasets, verifying the effectiveness of our method against 13 representative attacks and its resistance to potential adaptive attacks.

Parameter-oriented Scaling Consistency

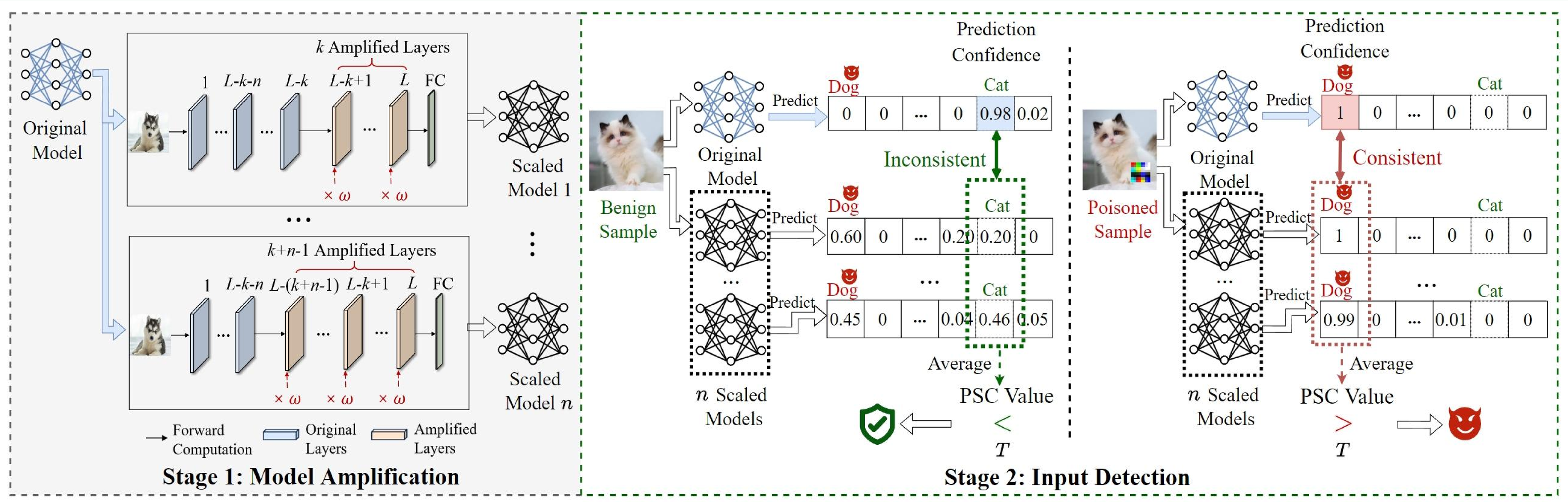
1. The average prediction confidence of the benign samples decreases during the parameter-amplified process.
2. The average prediction confidence of the poisoned samples remains nearly unchanged.



(a) Benign (b) BadNets (c) WaNet (d) BATT

Proposed Method

➢The Overview of IBD-PSC



➢IBD-PSC Consists of Two Steps:

1. **Model Amplification**

a) **Amplifiedm Model Parameters:**

$$\hat{\mathcal{F}}_k^{\omega} = \mathrm{FC} \circ \hat{f}_L^{\omega} \circ \hat{f}_{L-1}^{\omega} \circ \ldots \circ \hat{f}_{L-k+1}^{\omega} \circ \ldots \circ f_2 \circ f_1$$

b) **Layer Selection:**

$$\eta = \frac{1}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_r} \mathbb{I}\left(\operatorname{argmax}\left(\hat{\mathcal{F}}_k^{\omega}(\boldsymbol{x})\right) \neq y\right)$$

2. **Input Detection**

$$\mathrm{PSC}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=k}^{k+n-1} \hat{\mathcal{F}}_i^{\omega}(\boldsymbol{x})_{y'},$$

If $PSC(x) > T$, x is marked as a poisoned image

Experiments

Main defense results

Table 1. The performance (AUROC, F1) on the CIFAR-10 dataset. We mark the best result in boldface and failed cases (< 0.7) in red.

Attacks→	BadNets		Blend		PhysicalBA		IAD		WaNet		ISSBA		BATT		Avg.	
Defenses↓	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
STRIP	0.931	0.842	0.453	0.114	0.884	0.882	0.962	0.907	0.469	0.125	0.364	0.526	0.449	0.258	0.663	0.494
TeCo	0.998	0.970	0.675	0.678	0.748	0.689	0.909	0.920	0.923	0.915	0.901	0.942	0.914	0.673	0.858	0.834
SCALE-UP	0.962	0.913	0.644	0.453	0.969	0.715	0.967	0.869	0.672	0.529	0.942	0.894	0.959	0.911	0.731	0.757
IBD-PSC	**1.000**	**0.967**	**0.998**	**0.960**	**0.972**	**0.942**	**0.983**	**0.952**	**0.984**	**0.956**	**1.000**	**0.986**	**0.999**	**0.966**	**0.992**	**0.961**

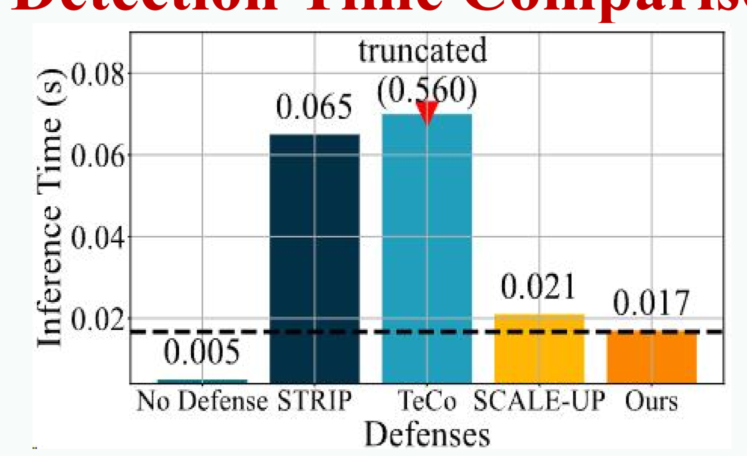
Table 2. The performance (AUROC, F1) on the GTSRB dataset. We mark the best result in boldface and failed cases (< 0.7) in red.

Attacks→	BadNets		Blend		PhysicalBA		IAD		WaNet		ISSBA		BATT		Avg.	
Defenses↓	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
STRIP	0.962	0.915	0.426	0.088	0.700	0.479	0.855	0.890	0.356	0.201	0.640	0.625	0.648	0.368	0.657	0.588
TeCo	0.879	0.905	0.917	0.913	0.860	0.673	0.955	0.962	0.954	0.935	0.941	0.947	0.829	0.673	0.907	0.858
SCALE-UP	0.913	0.858	0.579	0.421	0.762	0.709	0.885	0.860	0.309	0.149	0.733	0.691	0.902	0.876	0.700	0.669
IBD-PSC	**0.968**	**0.965**	**0.953**	**0.928**	**0.940**	**0.946**	**0.970**	**0.971**	**0.986**	**0.973**	**0.972**	**0.971**	**0.969**	**0.968**	**0.969**	**0.962**

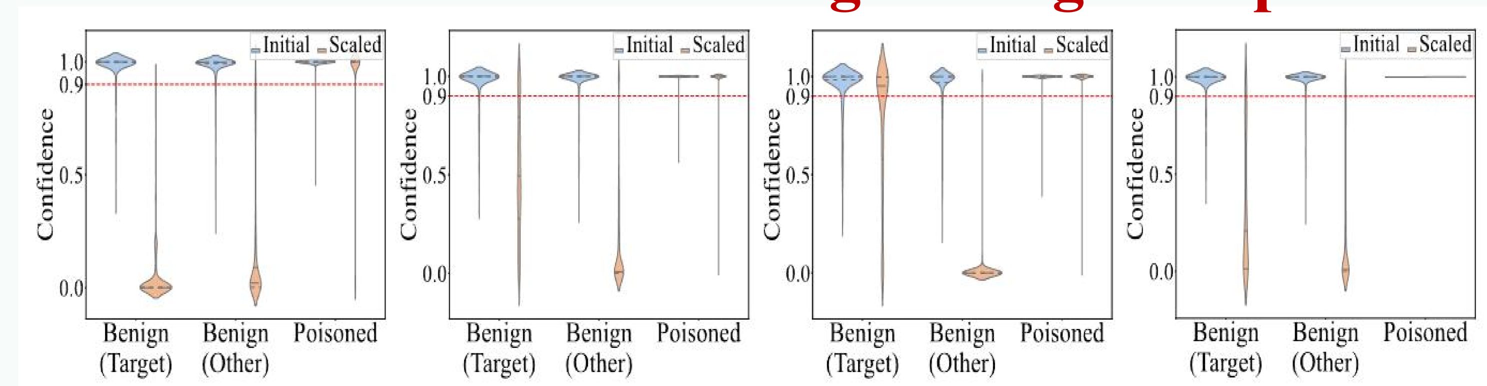
Table 3. The performance (AUROC, F1) on SubImageNet-200. We mark the best result in boldface and failed cases (< 0.7) in red.

Attacks→	BadNets		Blend		PhysicalBA		IAD		WaNet		ISSBA		BATT		Avg.	
Defenses↓	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
STRIP	0.840	0.828	0.799	0.772	0.618	0.468	0.528	0.419	0.563	0.356	0.768	0.765	0.554	0.361	0.681	0.596
TeCo	0.978	0.880	0.958	0.849	0.926	0.842	0.927	0.920	0.903	0.747	0.945	0.921	0.690	0.692	0.908	0.846
SCALE-UP	0.967	0.895	0.531	0.356	0.932	0.876	0.322	0.030	0.563	0.356	0.945	0.912	0.967	0.921	0.725	0.651
IBD-PSC	**1.000**	**0.992**	**0.989**	**0.833**	**0.994**	**0.988**	**0.994**	**0.996**	**0.967**	**0.981**	**0.989**	**0.987**	**0.998**	**0.998**	**0.990**	**0.974**

Detection Time Comparison

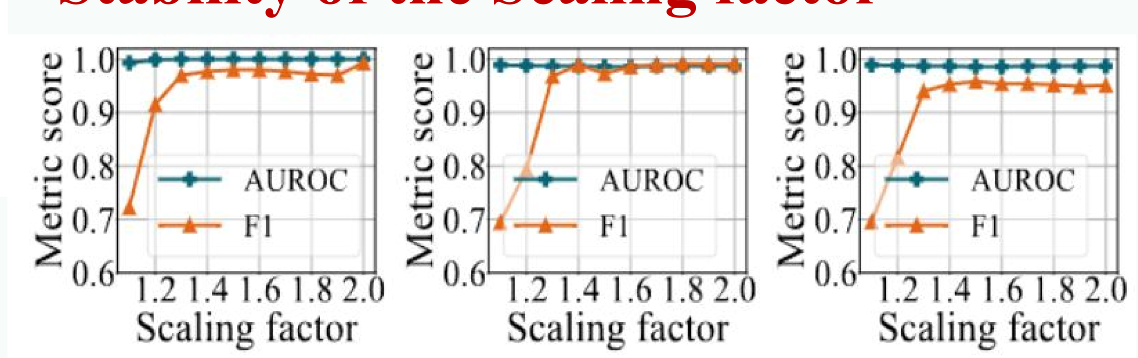


Performance on the target benign samples



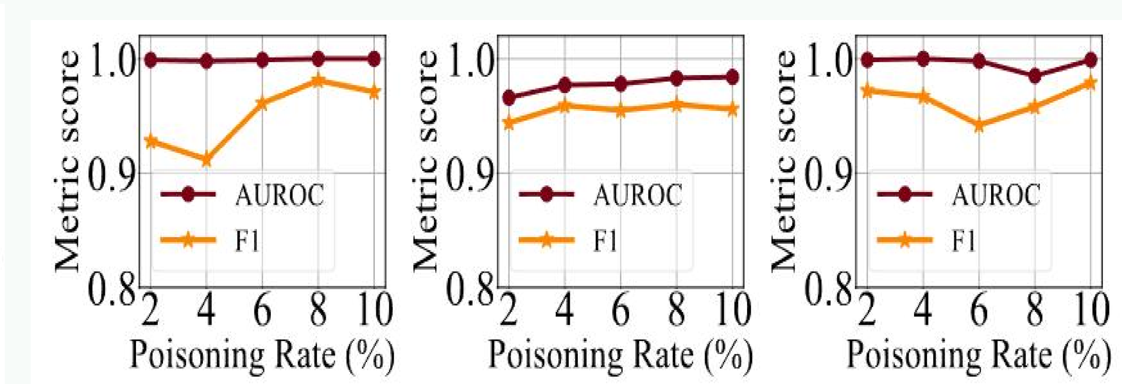
(a) BadNets (b) Blend (c) BATT (f) Ada-patch

Stability of the Scaling factor



(a) BadNets (b) WaNet (c) BATT

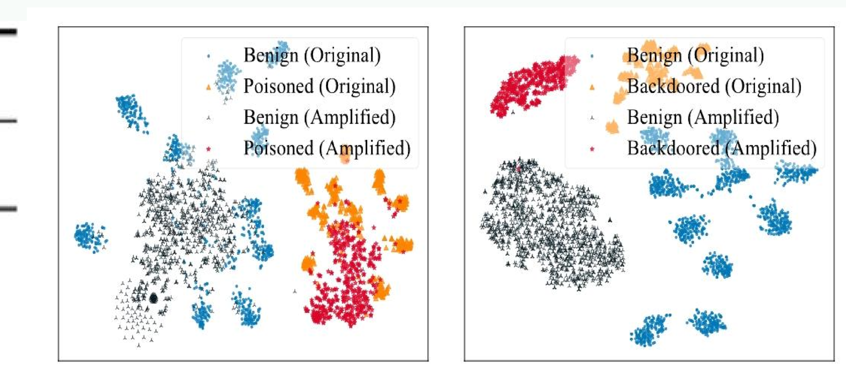
Robustness to the Poisoning Rates



Robustness against Adaptive Attacks

α→	0.2		0.5		0.9		0.99	
Attacks↓	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
BadNets	0.992	0.978	0.986	0.964	0.995	0.962	0.996	0.951
WaNet	0.947	0.949	0.956	0.942	0.931	0.927	0.819	0.862
BATT	0.986	0.968	0.994	0.956	0.982	0.975	0.979	0.959

Futher Understanding



(a) SCALE-UP (b) Ours