Table 1. Comparison between EX-RAY against other defenses on composite and reflection attacks

		TP	FP	FN	TN	Acc
Meta-Neural Analysis	Composite	15	6	5	14	0.73
	Reflection	11	8	9	12	0.58
DeepInspect	Composite	20	19	0	1	0.53
	Reflection	20	20	0	0	0.5
NeuronInspect	Composite	4	0	16	20	0.6
	Reflection	2	0	18	20	0.55
TABOR	Composite	3	0	17	20	0.58
	Reflection	2	0	18	20	0.55
ABS+EX-RAY	Composite	17	3	3	17	0.85
	Reflection	18	4	2	16	0.85

A. Experiment details

RBQ2: Comparison with baselines. Although EX-RAY is designed as an add-on, we agree that comparison of end-2-end pipelines is important too.

We hence add comparison with Meta neural analysis [13] on composite and reflection attacks. For black box backdoor defense [2], we do not find its code (even after reaching out to the authors). Besides, we also evaluate against three other SOTA defenses (DeepInspect [1], NeuronInspect [5] and TABOR [4]) on reflection and composite attacks as reviewer D suggests.

Table 1 compares the 4 SOTA defenses and ABS + EX-RAY. On CIFAR10, we use 20 benign models, 20 models trojaned with composite backdoor and 20 models trojaned with reflection backdoor. Meta neural analysis directly predicts whether a model is trojaned while the other three return a MAD score for each model. For DeepInspect, NeuronInspect and TABOR, we search the best possible bound of MAD score to separate trojaned and benign models. In Table 1, rows 2-9 show the results on composite and reflection attacks for Meta neural analysis, DeepInspect, NeuronInspect and TABOR. Rows 10-11 show the result of ABS + EX-RAY. We can see that ABS+EX-RAY outperfoms the SOTA methods, having at least 12% better accuracy on composite backdoors and 27% better on reflection backdoors.

During the TrojAI competition, performers tried many different SOTA methods [1, 3, 6, 7, 9–13] (including DeepInpect, Meta neural analysis and K-Arm). Except for K-Arm [9], all other methods perform worse than ABS + EX-RAY in rounds 2 to 4. K-Arm performs better than ABS + EX-RAY in round 3 but worse than ABS + EX-RAY in rounds 2 and 4.

RBQ3: Evaluation on SOTA attacks. We add comparison with WaNet [8] and input-aware dynamic attacks [8]. The results are in Table 2. On CIFAR10, we use 20 benign models, 20 models trojaned with WaNet and 20 models with input-ware backdoors. We set the bound for the trigger size to be 12.5% of the input. Our technique can achieve 82.5% accuracy on the former and 87.5% accuracy on the latter.

Table 2. ABS + EX-RAY input aware dynamic attacks and WaNet attacks

	TP	FP	FN	TN	Acc
Input-aware	17	2	3	18	0.875
Wanet	17	2	5	17	0.825

RDQ6: Ablation study of validation checks

The suggested ablation study's results on TrojAI rounds 2 to 4 are shown in Table 3. Observe that removing validation check results in 0.7% to 3% decrease in detection accuracy. These checks require masks computed by EX-RAY and cannot be incorporated to the vanilla ABS.

RDQ7: the second adaptive attack in the paper

We add an experiment for hyperparameter search of the weight of adaptive loss. The result is shown in Table 4. We think it may not be fair to compare the accuracy of adaptive attack model and that of an adversarial trained model. The former is not robust against adversarial examples. While users can bear the low accuracy for an adversarially trained model because it is robust, they may not be willing to use a model by the adaptive attack.

RDQ8: Another adaptive attack. We have conducted the suggested adaptive attack. In the adaptive attack, we first generate a trigger similar to a third class while having similar feature-level representations to the target class. We generate such triggers by optimizing two losses. The first is the cross entropy loss between the model output on images stamped with the trigger and the third class label (similar to adversarial noise for a third class). The second loss is the mean squared error loss between the inner activation of the images stamped with the trigger and the inner activation of the target class images (similar to adversarial feature-level attack). After generating the triggers, we use data poisoning to trojan the models. We do the experiment on CIFAR10. We choose label 0 as the target label and label 8 as the third label. We choose conv7 in NiN models as the feature layer and optimize neuron activations in this layer. We find that we need to enlarge the trigger size to have similar inner activations as the target label images. We generate triggers with 4 different sizes, 120, 140, 160, 200. The triggers are shown in Figure 1. We train 20 benign NiN models and 20 feature level adaptive attack NiN models for each trigger size.

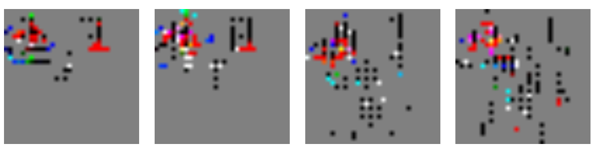
Table 5 shows the results of EX-RAY. Row 1 shows the different trigger sizes. Row 2 shows the mean squared activation differences. Observe that with the increase of trigger size, we can optimize the difference to a smaller value. A trigger with a small feature difference may be difficult to be detected. Rows 3 and 4 show the false positive and true

Table 3. EX-RAY w. and w.o. additional check; (T:276,C:552) means that there are 276 trojaned models and 552 clean models

	TrojAI R2						TrojAI R3						TrojAI R4					
	Polygon Trigger (T:276,C:552)			Filter Trigger (T:276,C:552)			Polygon Trigger (T:252,C:504)			Filter Trigger (T:252,C:504)			Polygon trigger (T:143,C:504)			Filter trigger (T:361,C:504)		
	TP	FP	Acc	TP	FP	Acc	TP	FP	Acc	TP	FP	Acc	TP	FP	Acc	TP	FP	Acc
W. additional check	198	19	0.883	204	32	0.874	200	46	0.870	149	39	0.812	105	53	0.859	242	46	0.809
W.o. additional check	206	33	0.876	216	71	0.844	207	71	0.843	158	62	0.793	110	77	0.829	275	93	0.793

Table 4. Adaptive attack two with more adaptive loss weights

Weight of adaptive loss	1	10	100	200	400	600	800	1000	10000
Acc (model/label)	0.89/0.73	0.88/0.73	0.87/0.7	0.87/0.7	0.86/0.69	0.845/0.66	0.84/0.66	0.82/0.64	0.1
ASR	0.99	0.99	0.99	0.98	0.94	0.98	0.96	0.97	-
FP/ # of clean models	0	0.2	0.2	0.2	0.35	0.45	0.6	0.65	-
TP/ # of clean models	1	1	1	1	1	1	1	1	-



(a) 120 (b) 140 (c) 160 (d) 200

Figure 1. Trigger for feature level trigger adaptive attack

positive rates. Observe that EX-RAY has 75% true positive rate when the trigger is 160 and 65% true positive rate when trigger size is 200. When the trigger size is 200, the trigger already covers a large part of the image. The attack becomes less meaningful.

Table 5. Feature level trigger adaptive attack

Trigger size	120	140	160	200
Mean squared feature difference	0.153	0.116	0.034	0.009
FP/ # of clean models	0.1	0.1	0.1	0.1
TP/ # of clean models	1	0.8	0.75	0.65

References

[1] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, pages 4658–4664, 2019. 2

[2] Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Black-box detection of backdoor attacks with limited information and data. *arXiv preprint arXiv:2103.13127*, 2021. 2

[3] N Benjamin Erichson, Dane Taylor, Qixuan Wu, and Michael W Mahoney. Noise-response analysis for rapid detection of backdoors in deep neural networks. *arXiv preprint arXiv:2008.00123*, 2020. 2

[4] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. *arXiv preprint arXiv:1908.01763*, 2019. 2

[5] Xijie Huang et al. Neuroninspect: Detecting backdoors in neural networks via output explanations. *arXiv preprint arXiv:1911.07399*, 2019. 2

[6] Susmit Jha, Sunny Raj, Steven Fernandes, Sumit K Jha, Somesh Jha, Brian Jalaian, Gunjan Verma, and Ananthram Swami. Attribution-based confidence metric for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 11826–11837, 2019. 2

[7] 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

[8] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 2

[9] Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, and Xiangyu Zhang. Backdoor scanning for deep neural networks through k-arm optimization. 2021. 2

[10] Karan Sikka, Indranil Sur, Susmit Jha, Anirban Roy, and Ajay Divakaran. Detecting trojaned dnns using counterfactual attributions. *arXiv preprint arXiv:2012.02275*, 2020. 2

[11] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks. In *27th {USENIX} Security Symposium*, 2018. 2

[12] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021. 2

[13] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. *arXiv preprint arXiv:1910.03137*, 2019. 2