A APPENDIX

A.1 ALTERNATIVE METRICS

Here we first examine whether the gradient magnitude of a neuron can be served as a good metric to distinguish between good and bad neurons and Figure 4 shows results. Similar to AM, solely employing gradient magnitude is hard to detect bad neurons.

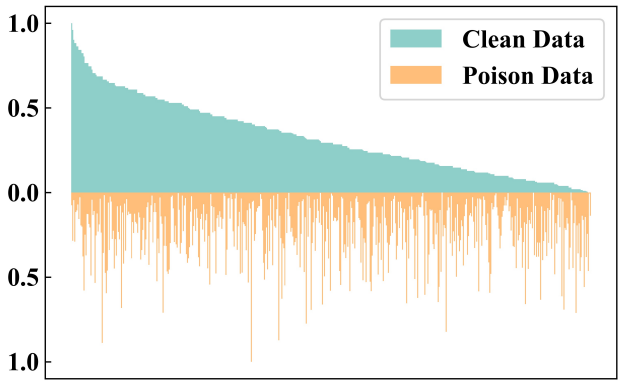

Figure 4: The gradient magnitude of each neuron with respect to clean data and poison data respectively. The evaluation settings follow Figure 3.

BS vs. Neuron Shapley & Integrated Gradient. Some literature (Sundararajan et al., 2017; Ghorbani & Zou, 2020) developed various metrics to quantify the importance of neurons. To show better practicality of BS, we compare our method with other importance evaluation metrics. Specifically, we select two common-used metrics: Neuron Shapley (Ghorbani & Zou, 2020) and Integrated Gradient Sundararajan et al. (2017). Table 3 reports the performance and overhead of three metrics. As can be seen, BS obtains competitive defense performance and only needs considerably smaller overheads compared with Integrated Gradient and Shapley value.

Table 3: The performance and overhead of different metrics against BadNet.

Metric	ACC	ASR	Time
BS	83.2	5.03	1.00
Integrated Gradient	83.25	5.10	10.76
Shapley Value	82.89	4.92	968.17

Table 4: The performance of AI-Lancet and WIPER against four backdoor attacks.

Defense	AI-Lancet		WIPER	
Attack	ACC	ASR	ACC	ASR
BadNet	80.15	5.21	**83.20**	**5.03**
BA	73.81	27.17	**82.82**	**5.22**
ETA	68.67	29.93	**82.41**	**5.80**
SSA	68.47	22.14	**81.7**	**4.76**

A.2 ADDITIONAL EXPERIMENT ON DIFFERENT DATASETS

Firstly, we demonstrate that the poor defense performance of AI-Lancet Zhao et al. (2021) against state-of-the-art attacks. Table 4 shows that AI-Lancet is vulnerable against BA, ETA, and SSA, which adopt more threatening triggers covering the entire images or being dynamic.

Attack	BadNet		BA		ETA		IA		SIG		TrojanNN	
Purified Layer	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
FC only	**83.20**	5.03	**82.82**	5.22	**82.41**	5.80	**82.04**	6.02	**82.41**	3.73	**84.67**	4.68
FC + 1 Conv	82.19	5.15	81.29	5.93	81.37	5.77	80.46	5.44	81.26	3.98	82.89	4.24
FC + 2 Conv	80.29	4.79	79.31	5.24	79.42	5.38	77.66	5.69	79.82	4.05	81.00	**4.01**
FC + 3 Conv	77.03	4.83	77.84	**4.34**	77.65	4.60	76.71	5.62	77.96	3.45	79.72	5.36
FC + 4 Conv	74.49	**4.51**	74.86	4.35	71.56	**4.04**	73.33	**4.96**	74.56	**3.22**	72.73	4.67

Table 5: The performance of WIPER when purifying different layers. Here n Conv indicates purifying the last n convolution layers.

Attack	BadNet		BA		ETA		IA		SIG		TrojanNN	
α	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
0.001	**83.57**	77.03	**84.09**	86.74	83.69	66.43	**84.07**	76.82	**83.57**	28.09	86.09	33.18
0.005	83.55	21.01	84.14	26.20	**84.41**	16.54	83.44	25.97	82.64	6.69	**86.22**	5.57
0.01	83.20	**5.03**	82.82	5.22	82.41	5.80	82.04	6.02	82.41	3.73	84.67	4.68
0.05	82.76	5.48	82.89	5.42	82.10	5.48	81.97	5.69	81.78	4.02	84.25	4.86
0.1	82.20	5.30	82.18	**5.07**	81.55	**3.96**	81.33	**4.72**	81.20	**3.44**	84.39	**4.39**

Table 6: The performance of WIPER with different α against six backdoor attacks in CIFAR-10.

Here we report some additional experimental results in SVHN and CIFAR-100. From experiments, we can further consolidate our conclusion that WIPER is an effective defense against backdoor attacks (also discussed in Section 5).

Figure 6 and Figure 7 illusrate the defense performance of different defenses on SVHN and CIFAR-100. Besides, Figure 9 and Figure 10 show the defense performance of WIPER with varying regularization items on SVHN and CIFAR-100. Table 9 and Table 10 report the defense performance of WIPER with different purifying strategies and neuron importance evaluation metrics on SVHN and CIFAR-100.

Defense	Before		Fine-tuning		KD		NAD		Fine-Pruning		IBAU		WIPER	
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNet	94.68	100	85.74	3.17	60.16	4.42	86.22	2.71	93.17	69.57	89.70	6.83	**94.14**	**0.49**
BA	94.62	100	86.54	1.44	52.4	2.61	86.47	3.11	92.29	79.65	87.95	9.91	**94.09**	**1.43**
ETA	94.59	99.98	87.61	6.34	53.08	4.98	87.56	6.28	93.41	68.52	91.22	10.98	**94.11**	**4.94**
IA	94.68	100	85.17	3.26	52.45	4.51	86.25	**1.87**	94.23	87.53	89.71	9.45	**94.53**	2.48
SIG	94.38	96.97	84.85	1.31	53.22	6.87	85.3	3.22	93.16	38.53	88.59	2.28	**93.69**	**1.31**
TrojanNN	94.71	100	88.75	8.42	49.64	4.81	87.45	31.97	92.31	97.82	88.22	11.34	**94.52**	**2.41**
IMC	94.15	100	87.52	10.56	52.62	5.15	86.41	32.53	93.50	98.23	87.52	18.16	**93.90**	**1.92**
WaveNet	94.13	99.36	84.68	2.81	55.86	6.26	84.84	4.57	92.54	40.86	88.06	3.25	**93.90**	**2.70**
SSA	94.60	97.41	85.22	2.22	52.83	7.59	85.30	4.07	94.95	40.76	87.52	5.27	**93.33**	**3.17**
LogitAttack	94.93	100	86.16	3.38	51.68	4.22	86.30	3.55	93.60	68.75	90.01	6.62	**93.51**	**2.14**

Table 7: The performance of WIPER compared with four state-of-the-art defense methods over six backdoor attacks in SVHN.

Defense	Before		Fine-tuning		KD		NAD		Fine-Pruning		IBAU		WIPER	
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNet	69.66	99.83	60.74	11.01	18.84	1.51	32.80	2.32	63.48	25.53	59.29	3.44	**63.71**	**1.12**
BA	68.68	99.81	61.68	23.32	15.32	1.59	33.24	2.46	63.08	26.68	61.62	9.08	**63.21**	**0.80**
ETA	68.98	98.43	62.71	17.99	17.52	1.44	31.03	3.11	**62.89**	37.63	60.82	1.67	62.35	**0.83**
IA	69.12	99.88	59.44	27.79	16.57	1.91	31.37	3.02	60.97	23.30	60.27	5.94	**62.11**	**0.55**
SIG	68.92	90.21	57.48	21.18	15.48	1.58	30.50	1.62	61.43	28.07	59.11	8.12	**63.89**	**0.26**
TrojanNN	70.84	94.66	61.97	56.76	16.61	1.57	30.38	32.03	61.45	78.99	60.58	11.56	**63.92**	**0.20**
IMC	69.08	99.55	62.41	62.15	17.52	1.98	31.19	35.40	61.62	87.98	60.47	15.16	**60.91**	**0.97**
WaveNet	67.68	98.84	59.49	22.07	16.00	1.29	31.08	2.46	62.04	27.87	59.95	8.37	**62.04**	**0.99**
SSA	68.03	97.76	58.80	22.00	15.50	1.38	30.45	0.21	62.23	28.98	61.00	6.46	**60.82**	**1.38**
LogitAttack	68.35	99.87	61.49	11.26	19.24	2.55	32.90	3.15	62.63	26.09	58.80	4.82	**62.00**	**0.53**

Table 8: The performance of WIPER compared with four state-of-the-art defense methods over six backdoor attacks in CIFAR-100.

A.3 Attack implementation details

We summarize the implementation details of ten backdoor attacks used in this paper as follows:

Attack	BadNet		BA		ETA		IA		SIG		TrojanNN	
Method	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
before	94.68	100.00	94.62	100.00	94.59	99.98	94.68	100.00	94.38	96.97	94.71	100.00
Pruning + AM	93.17	69.57	92.29	79.65	93.41	68.52	94.53	87.53	93.16	38.53	92.31	97.82
Pruning + BS	93.58	17.58	93.01	19.25	93.58	24.92	93.07	13.78	93.52	12.84	93.41	20.88
Purifying + AM	93.70	34.06	93.91	29.32	93.56	45.50	**94.57**	35.19	93.64	27.10	93.84	63.08
Purifying + BS	**94.14**	**0.49**	**94.09**	**2.87**	**94.11**	**4.94**	94.23	**9.48**	**93.69**	**1.48**	**94.52**	**2.41**

Table 9: Comparison of backdoor defense with different purifying strategies and neuron importance evaluation metrics on SVHN.

Attack	BadNet		BA		ETA		IA		SIG		TrojanNN	
Method	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
before	69.66	99.83	68.68	99.81	68.98	98.43	69.12	99.88	68.92	90.21	70.84	94.66
Pruning + AM	63.48	25.53	63.08	26.68	62.89	37.63	60.97	23.30	61.43	28.07	61.45	78.99
Pruning + BS	63.55	16.69	63.02	19.87	61.73	18.07	61.02	11.77	62.21	15.28	61.96	29.63
Purifying + AM	63.62	20.31	63.15	23.51	**63.01**	25.04	61.75	19.27	62.52	22.67	62.83	70.57
Purifying + BS	**63.71**	**1.12**	**63.21**	**0.80**	62.35	**0.83**	**62.11**	**0.55**	**63.89**	**0.26**	**63.92**	**0.20**

Table 10: Comparison of backdoor defense with different purifying strategies and neuron importance evaluation metrics on CIFAR100.

- **BadNet:** A 3×3 trigger with random pixel values is pasted in the top left corner of 5% training images, and labels tamper with 0.
- **Blend Attack:** We used the same trigger, a hello kitty image, in the original paper, blend ratio of 0.2, and inject rate of 0.05.
- **Enhanced Trigger Attack:** Following the original paper, we adopt the same random spatial transformation layer consisting of rotation and scale to pre-process the poison data. The associated parameters of the trigger used in this attack are identical to BadNet.
- **Invisible Attack:** We generated the same trigger with 32×32 resolution in the original paper and injection rate of 0.05.
- **Sinusoidal signal attack:** We used the backdoor trigger generation function in the original paper with $\delta = 20$ and $f = 6$ and injection rate of 0.1.
- **TrojanNN:** Based on the implementation of the original paper, we utilized the same reverse engineer technology to craft a 3×3 square trigger from the fully connected layer. Other parameters are set the same in BadNet.
- **IMC:** Similar to the original paper, we used a random noise with 3×3 size as the trigger and optimize the trigger during the training process.
- **WaveNet:** We used the corruption technique with a transformation probability of 0.2 defined in the original paper to process images.
- **SSA:** We added noises produced by the generator used in the original paper for 20% of images.
- **LogitAttack:** We poisoned 5% training data with the trigger of 3×3 similar to BadNet.

A.4 Defense implementation details

We list the detailed defense settings of Fine-Pruning, Fine-tuning, KD, NAD, and I-BAU for reference:

- **Fine-Pruning:** As suggested in the original paper, we pruned the last layer of the model with a pruning rate of 0.1, where AM is measured over a random subset of the training set as same to BS.
- **Fine-tuning:** Following the original paper, we adopted a standard fine-tuning procedure with a fixed learning rate of 0.1, momentum factor of 0.9, L_2 weight decay factor of 1×10^{-4}, and cross-entropy loss function to recover the backdoored model.

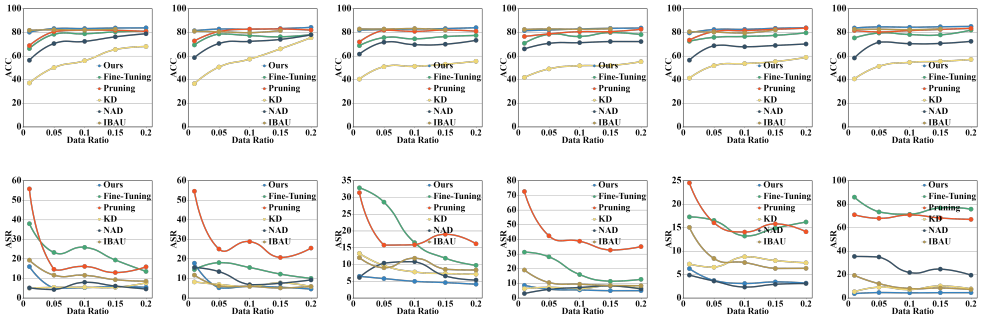

Figure 5: The defense performance of five defense methods over different data ratio (0.01, 0.05, 0.1, 0.15, 0.2) in CIFAR-10. The images from left to right indicate against BadNet, BA, ETA, IA, SIG, and TrojanNN.

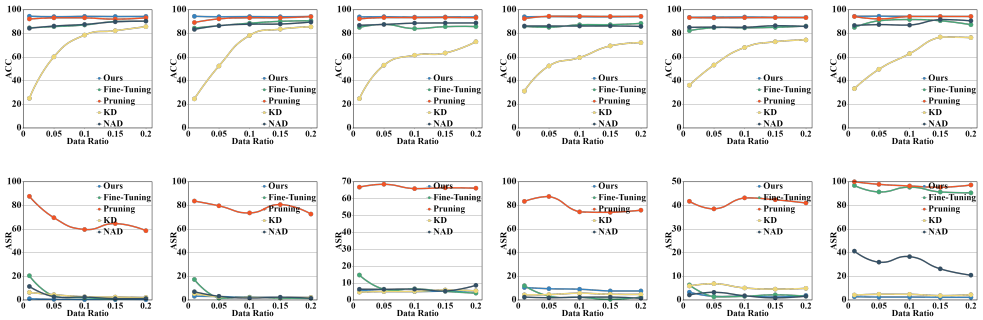

Figure 6: The defense performance of five defense methods over different data ratio (0.01, 0.05, 0.1, 0.15, 0.2) in SVHN.

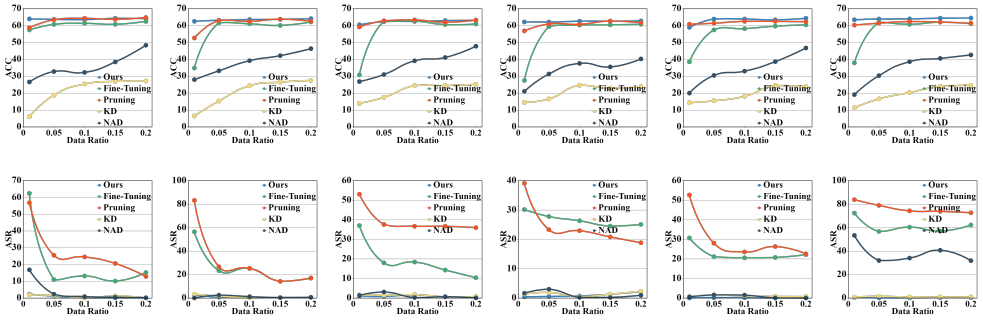

Figure 7: The defense performance of five defense methods over different data ratio (0.01, 0.05, 0.1, 0.15, 0.2) in CIFAR-100.

- **KD:** The standard knowledge approach is employed to recover the backdoored model, where cross-entropy loss and KL divergence are fused together as the overall loss function. The weights of the two loss terms are set to 0.5 along with the temperature of 2 and epoch of 50.
- **NAD:** We replicated the defense of NAD, where the backdoored model is fine-tuned for 10 epochs with an initial learning rate of 0.1 and a momentum of 0.9. The learning rate is divided by 10 after every 2 epochs.
- **I-BAU:** For I-BAU, we reused their original hyperparameter $C_\delta = 10$.

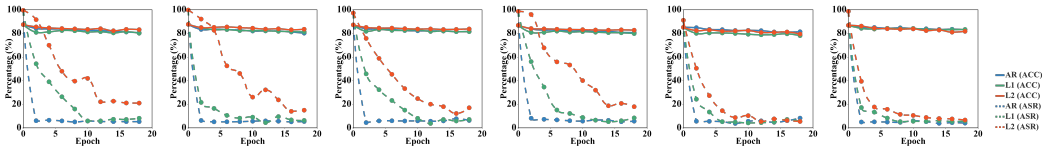

Figure 8: The performance of WIPER with different regularization items (L_1, L_2, and AR) on CIFAR-10. The images from left to right indicate against BadNet, BA, ETA, IA, SIG, and TrojanNN.

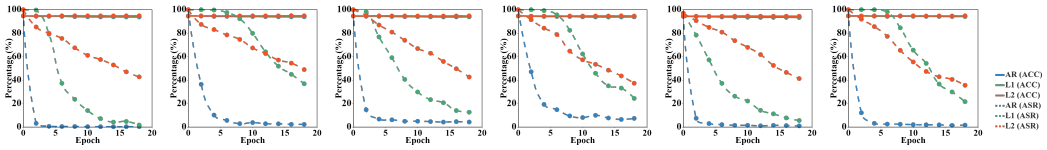

Figure 9: The performance of WIPER with different regularization items (L_1, L_2, and AR) on SVHN.

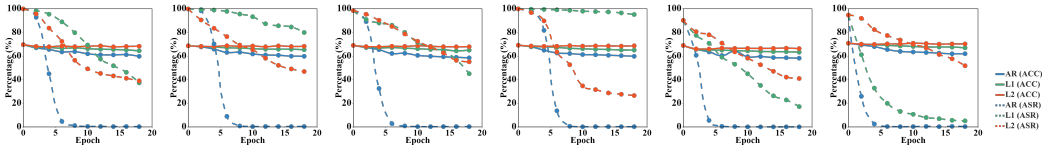

Figure 10: The performance of WIPER with different regularization items (L_1, L_2, and AR) on CIFAR-100.