Supplementary Material

Overview

The supplementary material is organized as follows.

- In Section A, we present the detailed derivation of the optimization algorithm of FT-SAM.
- In Section B, we introduce the details of the datasets, the implementation details of state-of-the-art (SOTA) attacks, SOTA defenses, and FT-SAM.
- In Section C, we display the defense results in comparison to SOTA defenses on different datasets and networks.
- In Section D, we show the defense effect of FT-SAM under different poisoning ratios compared to SOTA defenses.
- In Section E, we provide ablation study of the effectiveness of $\mathbf{T}_{\boldsymbol{w}}$ in FT-SAM.
- In Section F, we exhibit visualization analysis of the defense performance.

A. More Algorithmic Details on The Proposed Method

We provide a detailed derivation of the optimization problem in Section 3 of the main script here. The constraint optimization problem is defined as follows:

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{\epsilon} \in \mathcal{S}} \mathcal{L}(\boldsymbol{w}+\boldsymbol{\epsilon}), \tag{1}$$

where $\mathcal{L}(\boldsymbol{w}+\boldsymbol{\epsilon}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_{benign}}[\ell(f_{\boldsymbol{w}+\boldsymbol{\epsilon}}(\boldsymbol{x}),\boldsymbol{y})]$ with cross-entropy loss ℓ, $\mathcal{S} = \{\boldsymbol{\epsilon} : \|\mathbf{T}_{\boldsymbol{w}}^{-1}\boldsymbol{\epsilon}\|_2 \leq \rho\}$, $\rho > 0$ is the hyper-parameter for the budget of weight perturbation, and $\mathbf{T}_{\boldsymbol{w}}$ is the diagonal matrix.

Optimization. The optimization is inspired by [8]. Problem (1) can be efficiently solved by alternatively updating $\boldsymbol{w}$ and $\boldsymbol{\epsilon}$, as follows:
Inner Maximization: Given model weight $\boldsymbol{w}_t$, the weight perturbation $\boldsymbol{\epsilon}$ could be updated by solving the following sub-problem:

$$\max_{\boldsymbol{\epsilon} \in \mathcal{S}} \mathcal{L}(\boldsymbol{w}_t+\boldsymbol{\epsilon}). \tag{2}$$

Define $\tilde{\boldsymbol{\epsilon}} = \mathbf{T}_{\boldsymbol{w}}^{-1}\boldsymbol{\epsilon}$. According to first-order Taylor expansion, the approximation of the solution to Problem (2) is

$$\begin{aligned}
\tilde{\boldsymbol{\epsilon}}_{t+1} &= \arg\max_{\|\tilde{\boldsymbol{\epsilon}}\|_2 \leq \rho} \mathcal{L}(\boldsymbol{w}_t + \mathbf{T}_{\boldsymbol{w}_t}\tilde{\boldsymbol{\epsilon}}) \\
&\approx \arg\max_{\boldsymbol{\epsilon} \in \mathcal{S}} \mathcal{L}(\boldsymbol{w}_t) + \tilde{\boldsymbol{\epsilon}}^\top \mathbf{T}_{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t) \\
&= \arg\max_{\boldsymbol{\epsilon} \in \mathcal{S}} \tilde{\boldsymbol{\epsilon}}^\top \mathbf{T}_{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t) \\
&= \rho \frac{\mathbf{T}_{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t)}{\|\mathbf{T}_{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t)\|_2}.
\end{aligned} \tag{3}$$

Thus the inner problem can be solved as:

$$\boldsymbol{\epsilon}_{t+1} = \mathbf{T}_{\boldsymbol{w}_t}\tilde{\boldsymbol{\epsilon}}_{t+1} = \rho \frac{\mathbf{T}_{\boldsymbol{w}_t}^2 \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t)}{\|\mathbf{T}_{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t)\|_2}. \tag{4}$$

Outer Minimization: Given $\boldsymbol{\epsilon}_{t+1}$, the model weight $\boldsymbol{w}$ can be updated by solving the following sub-problem:

$$\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w} + \boldsymbol{\epsilon}_{t+1}), \tag{5}$$

which can be optimized by stochastic gradient descent, i.e., $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t + \boldsymbol{\epsilon}_{t+1})$ where η is the learning rate.

B. More Implementation Details

Datasets. We evaluate our method on CIFAR-10 [7], Tiny ImageNet [9], and GTSRB [20] following BackdoorBench [23]. For details, CIFAR-10 contains 60,000 images from 10 classes, with 5000 images per class for training and 1000 images per class for testing. Each image has a size of 32×32. Tiny ImageNet is a subset of ImageNet, which contains 100,000 training samples and 10,000 testing samples over 200 classes. Each image has a size of 64×64. GTSRB contains 39209 and 12630 images for training and testing from 43 classes. Each image has a size of 32×32.

Models. We evaluate our method on PreAct-ResNet18 [5] and VGG19-BN [19] networks. We compare our method with SOTA defense methods on three datasets and the two networks with a 10% poisoning ratio and 5% clean samples for defense. To study the effectiveness of our method under different poisoning ratios, we compare with SOTA defense methods on CIFAR-10 dataset and PreAct-ResNet18 network on 5% and 1% poisoning ratios.

Attack Details. We present some details about the backdoor attacks here. For BadNets-A2O and BadNets-A2A [4], we patch a 3×3 white square in the lower right corner of the images for CIFAR-10 and GTSRB datasets, and 6×6 white square for Tiny ImageNet. For Blended [3], we blend the poisoned samples with a Hello-Ketty image and the blended ratio is 0.1.

Defense Details. The seven SOTA defense methods can be divided into two types based on what the defender is given. AC [2] and ABL [11] assume that the defender is given a poisoned dataset, while the remaining defense methods assume that the defender can acquire a subset of clean samples and a backdoored model. The learning rate for all methods is set to 0.01 for FT and FT-SAM, and the batch size is set to 256. The threshold for ANP [24] is set to 0.4 since we find that the recommended threshold 0.2 fails to remove backdoors. For FT, the training epochs are set to 100 for CIFAR-10 and Tiny ImageNet, and 50 for GTSRB dataset. All other settings are consistent with those in BackdoorBench [23].

Details of Proposed Method. The most crucial hyper-parameter in FT-SAM is the perturbation radius ρ. We set $\rho = 2$ for CIFAR-10 and $\rho = 8$ for Tiny ImageNet and GTSRB on PreAct-ResNet18. For VGG19-BN, ρ is set to 6 for all three datasets. The epochs are set to 100 for CIFAR-10 and Tiny ImageNet, and 50 for GTSRB dataset. When the adaptive perturbation $\mathbf{T}$ is not applied to $\boldsymbol{w}$, the perturbation budget should be small to maintain the clean accuracy, where it is set to 0.5 in this work. All the experiments are conducted using SGD with momentum 0.9 and weight decay $1\mathrm{e}{-4}$.

C. Defense Results in Comparison to SOTA Defenses

The defense performances of our method compared to the seven SOTA defense methods on the three datasets and two networks are displayed in Table 1 to Table 4. **Note** that among all defenses, the one with the best performance is indicated in **boldface**, and the value with underline denotes the second-best result. We also compare our *FM-SAM* with two latest defense methods, *i.e.*, CLP [27] and NGD [6]. Due to space limit, we display the defense performance on CIFAR-10 dataset in Table 5

As shown in these tables, all the defense methods fail to balance the performance on both the clean accuracy (ACC) and the attack success rate (ASR) in all the situations except for FT-SAM, which is robust across all the attacks, datasets, and backbones. The average defense effectiveness rating (DER) and ASR rank first among these defenses. Although the ACC of the proposed method has dropped slightly, it usually falls within 1% on average.

D. Defense Results under Different Poisoning Ratios

To show the robustness of our method, we also test the defense performance under 5% and 1% poisoning ratios. We find that tuning the hyperparameter ρ by a dynamic strategy often yields better performance, and the value is within $[2, 10]$. The defense results with 5% benign data on CIFAR-10 dataset and PreAct-ResNet18 are shown in Table 6 and Table 7. As shown in the tables, attack performance drops when the poisoning ratio is only 1% except for Trojan, which demonstrates the power of this attack. FT-SAM significantly outperforms the other defense methods, although, at low poisoning ratios, it performs worse than at high poisoning ratios.

Table 1. Comparison with the state-of-the-art defenses on **GTSRB** dataset with 5% benign data on PreAct-ResNet18 (%).

Attack	**Backdoored** ACC/ASR	**FT** ACC/ASR/DER	**FP** [13] ACC/ASR/DER	**NAD** [12] ACC/ASR/DER	**AC** [2] ACC/ASR/DER	**NC** [22] ACC/ASR/DER	**ANP** [24] ACC/ASR/DER	**ABL** [11] ACC/ASR/DER	**i-BAU** [25] ACC/ASR/DER	**FT-SAM(Ours)** ACC/ASR/DER
BadNets-A2O[4]	96.35/95.02	97.60/45.77/74.63	**98.12/0.00/97.51**	97.54/79.94/57.54	57.05/16.71/69.51	93.47/0.02/96.06	96.79/0.21/97.41	94.53/**0.00**/96.60	96.35/**0.00/97.51**	96.36/0.17/97.43
BadNets-A2A[4]	97.05/92.33	98.04/42.01/75.16	**98.11**/0.51/95.91	97.84/2.46/94.93	96.14/80.93/55.25	94.05/0.50/94.41	96.73/50.39/70.81	12.30/7.32/50.13	95.30/0.43/95.08	96.97/**0.36/95.95**
Blended[3]	97.97/99.67	98.07/94.09/52.79	**98.31**/56.79/71.44	97.76/95.90/51.78	96.86/99.36/49.60	88.04/**2.61**/93.57	97.86/97.99/50.79	43.29/4.66/70.17	94.92/42.09/77.27	96.55/3.13/**97.56**
Input-aware[15]	97.17/97.09	97.58/47.14/74.97	97.98/1.36/97.86	97.47/65.94/65.57	38.43/51.69/43.33	95.24/1.16/97.00	96.20/1.12/97.50	9.97/59.83/25.03	96.03/1.13/97.41	**98.23/0.02/98.54**
LF[26]	97.97/99.58	98.00/83.83/57.88	97.87/69.19/65.15	**98.24**/79.76/59.91	36.25/98.80/19.53	92.22/0.18/96.82	98.03/60.36/69.61	26.29/0.68/63.61	88.69/7.43/91.44	96.52/**0.11/99.01**
SSBA[10]	98.31/99.77	98.39/98.88/50.45	**98.47**/60.19/69.79	98.37/96.95/51.41	53.59/80.78/37.14	90.75/1.51/95.35	98.36/98.98/50.39	50.89/0.50/75.92	87.27/**0.18**/94.27	95.99/0.70/**98.37**
Trojan[14]	98.33/100.00	**98.38**/87.72/56.14	98.00/42.08/78.80	98.01/0.10/**99.79**	96.90/100.00/49.28	92.29/0.02/96.97	98.17/86.92/56.46	89.65/**0.00**/95.66	93.66/**0.00**/97.66	96.92/0.11/99.24
Wanet[16]	95.71/98.20	98.69/0.02/99.09	**98.88**/0.28/98.96	98.32/0.04/99.08	61.67/2.14/81.01	96.34/0.01/99.09	97.42/0.18/99.01	40.36/86.25/28.30	97.50/0.26/98.97	98.61/**0.00/99.10**
Avg	97.35/97.71	98.09/62.43/67.64	**98.22**/28.80/84.43	97.94/52.64/72.50	67.11/66.30/50.58	92.80/0.75/96.16	97.45/49.52/74.00	45.91/19.91/63.18	93.72/6.44/93.70	97.02/**0.57/98.15**

Table 2. Comparison with the SOTA defenses on **CIFAR-10** dataset with 5% benign data on VGG19-BN (%).

Attack	**Backdoored** ACC/ASR	**FT** ACC/ASR/DER	**FP** [13] ACC/ASR/DER	**NAD** [12] ACC/ASR/DER	**AC** [2] ACC/ASR/DER	**NC** [22] ACC/ASR/DER	**ANP** [24] ACC/ASR/DER	**ABL** [11] ACC/ASR/DER	**i-BAU** [25] ACC/ASR/DER	**FT-SAM(Ours)** ACC/ASR/DER
BadNets-A2O[4]	90.42/94.43	89.06/2.34/95.36	89.11/12.39/90.37	86.80/5.77/92.52	84.79/93.01/47.90	88.97/5.63/93.68	**90.44**/87.64/53.39	80.30/23.23/80.54	87.69/3.13/94.29	89.02/**1.52/95.76**
BadNets-A2A[4]	91.16/84.39	89.65/**1.09/90.90**	89.70/1.91/90.51	88.15/1.60/89.89	85.85/88.82/47.35	91.16/84.39/50.00	**91.29**/81.87/51.26	20.05/14.90/49.19	86.86/2.19/88.95	89.58/1.22/90.80
Blended[3]	91.60/96.68	89.66/56.21/69.26	89.54/72.33/61.14	88.06/69.22/61.96	86.72/99.98/47.56	89.59/57.57/68.55	**91.49**/91.04/52.76	10.00/**0.00**/57.54	87.17/9.22/91.51	88.39/1.43/**96.02**
Input-aware[15]	88.66/94.58	**91.34**/19.54/87.52	**91.34**/5.42/94.58	91.00/14.11/90.23	48.01/22.54/65.69	91.30/4.39/95.09	89.67/20.43/87.07	30.10/99.66/20.72	88.30/3.70/95.26	90.59/**3.41/95.58**
CLA[18]	83.37/99.83	88.56/8.42/95.71	**88.80**/15.34/92.24	87.39/7.83/96.00	78.91/97.16/49.11	83.37/99.83/50.00	83.24/57.31/71.20	10.00/100.00/13.32	85.68/11.23/94.30	**88.80/7.50/96.17**
LF[26]	83.28/13.83	88.81/1.31/56.26	88.18/1.29/56.27	85.08/3.07/55.38	80.20/11.26/49.75	88.33/1.22/**56.31**	**89.20**/1.34/56.24	55.30/**0.14**/42.85	83.06/6.66/53.48	88.45/1.79/56.02
SIG[1]	83.48/98.87	88.11/2.90/97.98	**88.66**/8.28/95.29	86.14/6.30/96.28	78.84/99.52/47.68	83.48/98.87/50.00	82.94/**0.00/99.16**	10.00/**0.00**/62.69	84.50/4.47/97.20	88.59/2.00/**98.43**
SSBA[10]	90.85/95.11	89.07/62.26/65.54	89.26/65.33/64.09	88.11/52.22/70.07	85.81/90.63/49.72	90.85/95.11/50.00	**91.11**/76.00/59.56	10.00/**0.00**/57.13	85.61/12.37/88.75	89.25/3.30/**95.11**
Trojan[14]	91.57/100.00	90.30/6.63/96.05	90.04/29.71/84.38	87.01/5.17/95.14	86.02/1.64/**96.41**	**91.57**/100.00/50.00	89.27/**0.00/98.85**	10.00/100.00/9.22	86.40/2.69/96.07	88.14/5.10/95.74
Wanet[16]	84.58/96.49	**91.45**/2.79/96.85	91.10/3.36/96.57	90.68/10.23/93.13	85.51/83.73/56.38	84.58/96.49/50.00	89.82/**0.96/97.77**	10.00/100.00/12.71	89.61/2.40/97.05	91.36/1.00/**97.75**
Avg	87.90/87.42	**89.60**/16.35/85.14	89.57/21.54/82.54	87.84/17.55/84.06	80.07/68.83/55.75	88.32/64.35/61.36	88.85/41.66/72.73	24.58/43.79/40.59	86.49/5.81/89.69	89.22/**2.83/91.74**

Table 3. Comparison with the SOTA defenses on **Tiny ImageNet** dataset with 5% benign data on VGG19-BN (%).

Attack	**Backdoored** ACC/ASR	**FT** ACC/ASR/DER	**FP** [13] ACC/ASR/DER	**NAD** [12] ACC/ASR/DER	**AC** [2] ACC/ASR/DER	**NC** [22] ACC/ASR/DER	**ANP** [24] ACC/ASR/DER	**ABL** [11] ACC/ASR/DER	**i-BAU** [25] ACC/ASR/DER	**FT-SAM(Ours)** ACC/ASR/DER
BadNets-A2O[4]	43.56/99.96	49.84/99.45/50.26	49.49/96.74/51.61	49.35/0.27/99.84	43.04/99.99/49.74	43.57/99.96/50.00	43.42/4.46/97.68	41.10/**0.00**/98.75	45.02/98.97/50.49	**50.08**/0.14/**99.91**
BadNets-A2A[4]	54.44/50.74	53.97/49.22/50.53	53.13/**1.33/74.05**	54.13/36.48/56.98	42.98/36.57/51.36	51.14/30.65/58.40	**54.40**/1.99/**74.36**	37.10/31.19/51.11	46.72/36.46/53.28	52.91/3.24/72.99
Blended[3]	50.68/97.08	50.04/80.81/57.81	49.78/64.10/66.04	50.24/57.45/69.59	41.26/96.10/45.78	48.84/**0.12/97.56**	**50.44**/95.46/50.69	40.84/12.14/87.55	45.57/89.55/51.21	49.05/6.01/94.72
Input-aware[15]	53.20/99.84	53.33/0.06/**99.89**	53.16/1.42/99.19	**53.50**/0.14/99.85	41.39/98.49/44.77	53.29/0.08/99.88	53.41/**0.01/99.91**	40.48/3.48/91.82	47.97/6.31/94.15	51.78/0.26/99.08
LF[26]	48.92/7.73	50.23/0.03/53.85	50.29/0.02/**53.85**	50.44/0.08/53.82	39.28/9.90/45.18	46.42/**0.01**/52.61	**50.68**/0.39/53.67	34.89/8.79/42.99	43.81/0.03/51.29	48.78/0.02/53.78
SSBA[10]	51.39/97.92	50.58/88.93/54.09	50.27/32.89/81.95	50.23/71.66/62.55	42.40/97.50/45.72	49.39/**0.05**/97.93	51.41/97.26/50.33	40.68/0.26/93.47	47.43/90.02/51.97	**51.49**/1.70/**98.11**
Trojan[14]	51.50/99.98	50.94/98.84/50.29	50.25/16.17/91.28	51.02/99.96/49.77	42.92/99.90/45.75	48.85/0.11/98.61	**51.57**/97.06/51.46	36.97/**0.00**/92.72	43.77/99.69/46.28	49.59/0.04/**99.01**
Wanet[16]	54.11/99.98	**54.21**/0.14/**99.92**	53.69/19.59/89.99	53.67/**0.10**/99.72	41.14/96.03/45.49	51.86/0.11/98.81	54.18/60.06/69.96	41.67/1.16/93.19	48.32/88.33/52.93	51.73/0.50/98.55
Avg	50.98/81.65	**51.64**/52.19/64.58	51.26/29.03/75.99	51.57/33.27/74.02	41.80/79.31/46.72	49.17/16.39/81.72	51.19/44.58/68.51	39.22/7.13/81.45	46.08/63.67/56.45	50.68/**1.49/89.52**

Table 4. Comparison with the SOTA defenses on **GTSRB** dataset with 5% benign data on VGG19-BN (%).

Attack	**Backdoored** ACC/ASR	**FT** ACC/ASR/DER	**FP** [13] ACC/ASR/DER	**NAD** [12] ACC/ASR/DER	**AC** [2] ACC/ASR/DER	**NC** [22] ACC/ASR/DER	**ANP** [24] ACC/ASR/DER	**ABL** [11] ACC/ASR/DER	**i-BAU** [25] ACC/ASR/DER	**FT-SAM(Ours)** ACC/ASR/DER
BadNets-A2O[4]	97.28/93.44	97.42/30.37/81.54	**97.63**/0.05/**96.69**	97.43/89.78/51.83	32.14/65.17/31.56	94.78/**0.00**/95.47	97.10/0.02/96.62	3.56/**0.00**/49.86	91.01/20.51/83.33	95.98/0.03/96.05
BadNets-A2A[4]	97.59/93.29	**98.40**/88.12/52.59	98.34/69.48/61.91	97.83/88.16/52.57	95.00/89.62/50.54	95.34/1.09/94.98	98.06/86.56/53.37	10.55/8.30/48.97	96.83/0.37/96.08	96.86/**0.20/96.18**
Blended[3]	97.06/99.12	97.43/97.21/50.96	**97.66**/97.30/50.91	97.28/96.83/51.15	95.31/98.38/49.50	94.95/56.62/70.20	97.13/98.56/50.28	3.56/**0.00**/52.81	96.38/63.49/67.48	97.18/1.73/**98.70**
Input-aware[15]	96.32/85.03	91.34/19.54/80.25	**97.65**/0.49/92.27	97.12/1.63/91.70	31.88/17.28/51.65	96.53/0.24/92.39	96.94/**0.00/92.51**	1.74/81.95/4.25	94.98/38.73/72.48	97.25/0.04/**92.49**
LF[26]	97.25/0.42	97.07/0.03/50.10	**97.59**/0.02/**50.20**	97.43/0.02/**50.20**	28.04/3.88/15.40	95.04/0.04/49.09	97.35/0.45/50.00	5.53/43.11/4.14	89.75/0.01/46.46	95.19/**0.00**/49.18
SSBA[10]	97.85/99.43	**98.00**/98.97/50.23	97.93/98.77/50.33	97.75/98.57/50.38	31.61/71.26/30.97	94.90/67.48/64.50	97.85/99.34/50.04	21.54/**0.00**/61.56	86.98/99.92/44.57	96.00/1.81/**97.88**
Trojan[14]	97.97/100.00	97.68/8.27/95.72	**98.00**/99.99/50.00	97.76/6.34/96.73	97.02/100.00/49.52	95.61/0.02/98.82	97.85/97.28/51.30	5.23/**0.00**/53.63	96.01/**0.00**/99.02	96.99/0.02/**99.51**
Wanet[16]	94.76/98.32	98.36/25.14/86.59	98.66/1.31/98.51	98.37/0.20/99.06	32.05/4.03/65.79	96.41/7.30/95.51	98.21/0.10/99.11	12.12/58.00/28.84	87.17/10.30/90.21	**98.76/0.04/99.14**
Avg	97.01/83.63	96.96/45.95/68.50	**97.93**/45.93/68.85	97.62/47.69/67.95	55.38/56.20/43.12	95.45/16.60/82.62	97.56/47.79/67.90	7.98/23.92/38.01	92.39/29.17/74.95	96.78/**0.48/91.14**

Table 5. Comparison with two latest defenses on CIFAR-10 dataset with 5% benign data on PreAct-ResNet18 (%)

Defense	ATTCK	BadNets-A2O[4]	Blended[3]	Input-aware[15]	CLA[18]	LF[26]	SIG[1]	SSBA[10]	Trojan[14]	Wanet[16]
CLP[27]	ACC	88.68	89.08	90.66	83.29	87.34	82.03	91.27	92.18	90.06
	ASR	83.64	97.49	11.97	0.00	99.02	99.58	12.68	99.99	2.09
NGD[6]	ACC	91.17	92.41	94.11	91.70	92.30	91.13	92.20	92.85	93.06
	ASR	2.30	52.01	0.91	5.43	89.33	1.39	77.19	29.29	2.03
FT-SAM(Ours)	ACC	92.21	92.44	93.76	90.72	91.07	91.16	92.12	92.75	92.87
	ASR	1.63	4.91	1.07	3.52	3.81	0.80	2.80	4.12	0.96

Table 6. Comparison with the SOTA defenses with a **5% poisoning ratio** on CIFAR-10 dataset with 5% benign data on PreAct-ResNet18 (%).

Attack	**Backdoored** ACC/ASR	**FT** ACC/ASR/DER	**FP** [13] ACC/ASR/DER	**NAD** [12] ACC/ASR/DER	**AC** [2] ACC/ASR/DER	**NC** [22] ACC/ASR/DER	**ANP** [24] ACC/ASR/DER	**ABL** [11] ACC/ASR/DER	**i-BAU** [25] ACC/ASR/DER	**FT-SAM(Ours)** ACC/ASR/DER
BadNets-A2O[4]	92.35/89.52	90.83/2.50/92.75	92.10/1.47/**93.90**	89.92/1.98/92.56	88.67/88.33/48.75	90.88/1.62/93.22	92.23/2.80/93.30	81.58/**0.00**/89.38	89.61/1.00/92.89	**92.27**/2.12/93.66
BadNets-A2A[4]	92.54/65.85	91.78/0.93/82.08	92.37/1.02/82.33	91.19/1.38/81.56	87.71/54.64/53.19	89.46/1.25/80.76	91.93/1.45/81.90	42.31/38.38/38.62	90.69/1.67/81.17	**92.54/0.91/82.47**
Blended[3]	93.66/94.82	93.18/83.63/55.35	93.10/9.64/92.31	93.08/66.46/63.89	89.27/87.52/51.46	93.00/87.53/53.31	**93.24**/82.26/56.07	73.23/**0.19**/87.10	86.73/1.30/**93.30**	91.07/8.27/91.98
Input-aware[15]	91.51/93.05	93.08/66.97/63.04	93.17/26.71/83.17	93.28/92.26/50.40	89.05/72.60/59.00	93.23/82.31/55.37	91.06/13.31/89.65	85.54/83.97/51.56	91.28/22.10/85.36	**93.69/6.23/93.41**
CLA[18]	93.47/99.33	92.67/96.29/51.12	92.38/39.00/79.62	92.38/90.19/54.03	89.87/96.14/49.79	**93.47**/99.33/50.00	92.76/23.16/87.73	73.52/99.67/40.03	88.26/40.60/76.76	92.86/**5.70/96.51**
LF[26]	93.51/97.29	**93.19**/96.23/50.37	92.11/69.07/63.41	92.93/94.96/50.88	89.12/95.33/48.78	93.04/54.28/71.27	93.01/73.98/61.41	61.19/94.11/35.43	89.85/28.73/82.45	92.74/**3.81/96.35**
SIG[1]	93.29/95.06	92.73/92.41/51.04	92.87/43.99/75.32	92.21/82.62/55.68	89.66/94.44/48.49	**93.29**/95.06/50.00	92.78/97.47/49.75	57.72/**0.00**/79.74	88.04/7.30/91.25	92.62/0.61/**96.89**
SSBA[10]	93.08/94.09	92.62/83.63/55.00	92.23/13.70/89.77	92.35/86.03/53.66	89.12/86.92/51.60	**93.08**/94.09/50.00	93.07/79.38/57.35	78.75/**0.94**/89.41	90.62/2.62/94.50	92.35/3.84/**94.76**
Trojan[14]	93.61/99.99	92.82/99.87/49.67	92.77/88.68/55.24	93.08/31.86/83.80	89.61/99.97/48.01	93.03/99.79/49.81	**93.26**/99.99/49.83	70.19/**0.00**/88.28	89.19/4.89/95.34	93.12/6.84/**96.33**
Wanet[16]	93.38/97.27	**93.45**/19.96/88.65	93.01/1.50/97.70	93.31/8.56/94.32	88.13/58.24/66.89	93.38/97.27/50.00	92.93/**0.31/98.26**	60.52/99.04/33.57	89.16/1.58/95.74	93.27/0.80/**98.18**
Avg	93.04/92.63	92.64/64.24/63.91	92.61/29.48/81.28	92.37/55.63/68.08	89.02/83.41/52.60	92.59/71.25/60.37	92.63/47.41/72.52	68.46/41.63/63.31	89.34/11.18/88.88	**92.65/3.91/94.05**

E. Ablation Study of The Effectiveness of $\mathbf{T}_{\boldsymbol{w}}$

In this section, we first study the effectiveness of the adaptive constraint $\mathbf{T}_{\boldsymbol{w}}$, then we show the experimental results on the defense that directly regularizes the l_2 weight norm.

Effectiveness of The Adaptive Constraint $\mathbf{T}_{\boldsymbol{w}}$. The constraint without adaptive perturbation is equal to the situation where $\mathbf{T}_{\boldsymbol{w}}$ is set to an identity matrix. The comparison result is shown in Table 8. As shown in the table, the method without adaptive constraints has a lower ACC and a higher ASR on average. This gap is more pronounced when encountering complex attacks. It demonstrates the necessity of the adaptive constraint to the perturbation in FT-SAM.

Defense results of The l_2 weight norm regularization. To show the effectiveness of FT-SAM, we also test the defense performance by directly fine-tuning with regularizing the l_2 norm on the network parameters, *i.e.*, the loss function is

$$\min_{\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_{benign}} \left[\ell(f_{\boldsymbol{w}}(\boldsymbol{x}),\boldsymbol{y})\right] + \gamma\|\boldsymbol{w}\|_2^2, \tag{6}$$

where $\gamma > 0$ is the hyper-parameter. We test this method on several complex attacks and the results under different values of $\gamma > 0$ are shown in Table 9. It is observed that regularizing weights can also weaken backdoor attacks to a certain extent. However, the hyper-parameter γ is very sensitive to different attacks, and removing backdoors completely usually results in a large drop in clean accuracy. On the contrary, our method is more robust to different attacks, showing the effectiveness of our method on perturbing the backdoor-related weights.

F. Visualization Analysis

Visualization of Gradient Change within epochs. In order to conduct a more comprehensive investigation into the intricate connection between backdoor-related neurons and the gradient norms derived through FT-SAM computation, we visualize the gradient norms for each neuron situated in the last convolutional layer of the defense models. This visualization is performed across the first batch of each epoch, ranging from the first to the eighth epoch. To facilitate clarity within this visualization, the neurons are sorted by TAC calculated by the specific backdoored model instead of the changing defense model. The attack success rate (ASR) is also labeled in figures to better show the changes of gradient norms within different ASR. The last convolution layer is chosen since the backdoor removal effect of FT-SAM is layer-wisely accumulated, and thus weight norm

Table 7. Comparison with the SOTA defenses with a **1% poisoning ratio** on CIFAR-10 dataset with 5% benign data on PreAct-ResNet18 (%).

Attack	**Backdoored** ACC/ASR	**FT** ACC/ASR/DER	**FP** [13] ACC/ASR/DER	**NAD** [12] ACC/ASR/DER	**AC** [2] ACC/ASR/DER	**NC** [22] ACC/ASR/DER	**ANP** [24] ACC/ASR/DER	**ABL** [11] ACC/ASR/DER	**i-BAU** [25] ACC/ASR/DER	**FT-SAM(Ours)** ACC/ASR/DER
BadNets-A2O[4]	93.12/74.20	90.83/2.50/84.71	92.88/2.44/85.76	92.38/10.87/81.30	89.25/11.84/79.24	92.77/30.67/71.59	**93.09**/5.84/84.16	82.56/**0.83**/81.40	90.05/2.80/84.17	92.74/1.31/**86.25**
BadNets-A2A[4]	93.42/28.62	91.78/**0.93**/63.03	92.32/1.07/63.23	92.28/1.87/62.81	88.72/1.86/61.03	**93.42**/28.60/50.01	93.12/4.78/61.77	52.76/21.93/33.02	88.84/2.11/60.97	92.74/1.11/**63.42**
Blended[3]	93.69/73.88	93.18/83.63/49.75	92.99/5.97/83.61	93.24/47.94/62.74	89.62/41.51/64.15	**93.69**/73.88/50.00	93.25/49.33/62.05	74.45/24.90/64.87	86.56/7.26/79.75	91.85/**4.62/83.71**
Input-aware[15]	91.15/68.53	93.08/66.97/50.78	93.07/20.81/73.86	93.13/84.92/50.00	89.99/56.44/55.46	93.19/57.83/55.35	91.58/63.79/52.37	58.76/21.92/57.11	90.72/10.62/78.74	**93.31/1.47/83.53**
CLA[18]	93.71/94.41	92.67/96.29/49.48	93.03/30.77/81.48	93.31/87.73/53.14	89.58/11.34/89.47	93.38/91.00/51.54	**93.50**/93.83/50.18	65.80/14.00/76.25	88.67/11.01/89.18	92.96/**4.97/94.35**
LF[26]	93.29/85.94	**93.19**/96.23/49.95	92.31/61.76/61.60	92.91/77.48/54.04	89.38/78.67/51.68	91.09/**3.64**/90.05	93.08/45.53/70.10	56.17/63.32/42.75	85.67/72.01/53.16	92.10/4.50/**90.13**
SIG[1]	93.68/78.68	92.73/92.41/49.53	92.02/67.74/54.64	93.13/78.11/50.01	90.12/79.77/48.22	**93.68**/78.68/50.00	93.47/78.38/50.05	65.12/**0.00**/75.06	90.11/31.69/71.71	91.43/3.29/**86.57**
SSBA[10]	93.51/70.69	92.62/83.63/49.56	93.17/7.20/81.57	93.15/54.54/57.89	89.29/31.38/67.55	93.16/54.89/57.73	**93.28**/24.48/72.99	59.42/65.03/35.78	90.42/**1.10**/83.25	92.96/1.81/**84.16**
Trojan[14]	93.80/99.89	92.82/99.87/49.52	92.91/98.32/50.34	93.45/99.87/49.84	89.95/99.73/48.15	93.42/99.91/49.81	**93.51**/99.86/49.87	61.68/43.73/62.02	87.56/59.07/67.29	93.14/**8.23/95.50**
Wanet[16]	93.03/81.05	**93.45**/19.96/80.54	93.33/**0.49/90.28**	93.27/2.59/89.23	89.18/4.67/86.27	93.21/3.51/88.77	92.75/1.24/89.77	29.86/81.91/18.42	90.64/1.19/88.73	93.21/0.76/90.15
Avg	93.24/75.59	92.64/64.24/57.68	92.80/29.66/72.64	93.03/54.59/61.10	89.51/41.72/65.12	**93.10**/52.26/61.48	93.06/46.71/64.33	60.66/33.76/54.67	88.92/19.89/75.69	92.64/**3.21/85.78**

Table 8. Comparison with the state-of-the-art defenses on CIFAR-10 dataset with 5% benign data on PreAct-ResNet18 (%). The better result between the two is indicated in **boldface**.

Model	BadNets-A2O[4] ACC/ASR/DER	BadNets-A2A[4] ACC/ASR/DER	Blended[3] ACC/ASR/DER	Input-aware[15] ACC/ASR/DER	CLA[18] ACC/ASR/DER	LF[26] ACC/ASR/DER	SIG[1] ACC/ASR/DER	SSBA[10] ACC/ASR/DER	Trojan[14] ACC/ASR/DER	Wanet[16] ACC/ASR/DER	Avg ACC/ASR/DER
Backdoored	91.82/93.79/-	91.89/74.42/-	93.44/97.71/-	94.03/98.35/-	84.55/99.93/-	93.01/99.06/-	84.49/97.87/-	92.88/97.07/-	93.47/99.99/-	92.80/98.90/-	91.24/95.71/-
w/o Adaptive	90.85/**1.53**/95.64	90.95/1.39/86.05	91.30/**2.44/96.56**	92.94/1.39/97.93	89.95/6.19/96.87	90.77/6.73/95.04	89.84/**0.49/98.69**	90.74/5.78/94.57	90.80/14.02/91.65	91.94/1.89/98.07	91.01/4.19/95.65
w/ Adaptive	**92.21**/1.63/**96.08**	**91.87/1.03/86.69**	**92.44**/4.91/95.90	**93.76/1.07/98.51**	**90.72/3.52/98.21**	**91.07/3.81/96.65**	**91.16**/0.80/98.53	**92.12/2.80/96.75**	**92.75/4.12/97.57**	**92.87/0.96/98.97**	**92.10/2.47/96.62**

Table 9. Defense results of l_2 weight norm regularization on CIFAR-10 dataset with 5% benign data on PreAct-ResNet18 (%).

Attack γ	BadNets-A2O[4] ACC/ASR/DER	Blended[3] ACC/ASR/DER	Input-aware[15] ACC/ASR/DER	LF[26] ACC/ASR/DER	SSBA[10] ACC/ASR/DER	Trojan[14] ACC/ASR/DER
(Attack)	91.82/93.79/-	93.44/97.71/-	94.03/98.35/-	93.01/99.06/-	84.49/97.87/-	92.88/97.07/-
0.001	90.69/1.27/**95.70**	92.77/74.13/61.45	93.97/10.55/**89.84**	92.40/85.36/56.55	92.37/67.56/65.16	92.70/26.38/85.25
0.005	89.24/1.33/94.94	91.80/7.73/94.17	93.67/10.76/89.73	91.75/33.13/82.33	91.69/9.24/94.31	92.23/11.03/**92.69**
0.01	88.99/0.80/95.08	89.13/1.24/**96.08**	92.07/11.06/89.58	90.54/13.59/**91.50**	89.04/2.58/**97.64**	89.62/9.98/91.91
0.05	36.55/0.11/69.20	28.44/1.59/65.56	41.08/6.87/67.47	35.85/17.08/62.41	49.13/2.18/80.16	44.50/9.27/69.71
0.1	18.47/2.59/58.93	12.79/6.41/55.33	18.90/3.10/58.27	17.93/2.27/60.85	10.10/0.51/61.48	13.66/2.87/57.49

changes in the last layer after fine-tuning all layers are most obvious to show the performance. As shown in Figure 1, the norms of the gradient are obviously positively related to the backdoor-related neurons until the backdoor is removed. This observation serves to unveil the intrinsic mechanism underlying FT-SAM's framework.

Grad-CAM Visualization. Figure 2 to 5 show the defense effect of our method on BadNets, Blended, SIG, and Wanet attacks by Grad-CAM [17]. The top rows show the poisoned samples, while the second and third rows show the Grad-CAM figures on the backdoored models and the defense models, respectively. Figure 2 to 4 belong to visible backdoor attacks. Comparing the highlighted area of the heat maps of the backdoored models and defense models, the defense models concentrate on the subject region of the images instead of the trigger features. Figure 5 shows the invisible backdoor attack. The defense models focus more on the subject region, whereas the backdoored models show similar areas of interest in all these images.

T-SNE Visualization. We provide more T-SNE [21] visualization figures of our method as shown in Figure 6. compared to the first row which exhibits clustering of poisoned features in the feature space of the backdoored models, the proposed defense method successfully breaks up these poisoned features and makes them distribute around the normal features.

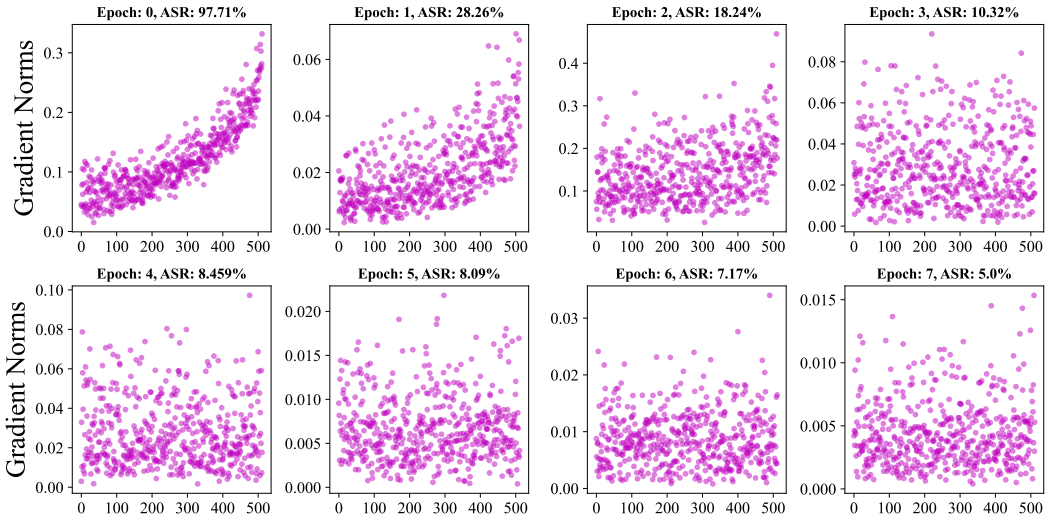

Figure 1. A comparison of the gradient norms for each neuron in the last convolution layer of the defense models, which are calculated during the first batch of each epoch. The neurons are sorted by TAC of the backdoored model.

Figure 2. Grad-CAM visualization of regions contributed to model decision under BadNets attack and FT-SAM defense with PreAct-ResNet18 on CIFAR-10.

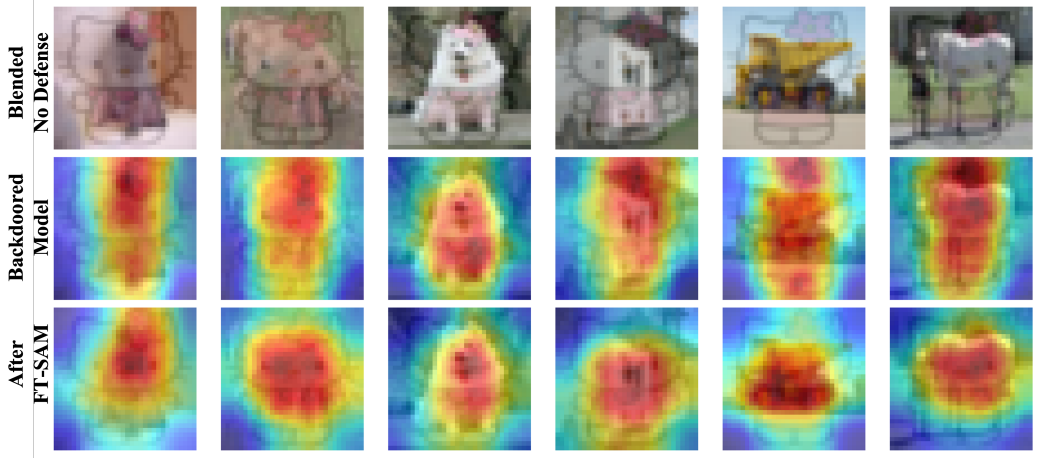

Figure 3. Grad-CAM visualization of regions contributed to model decision under Blended attack and FT-SAM defense with PreAct-ResNet18 on CIFAR-10.

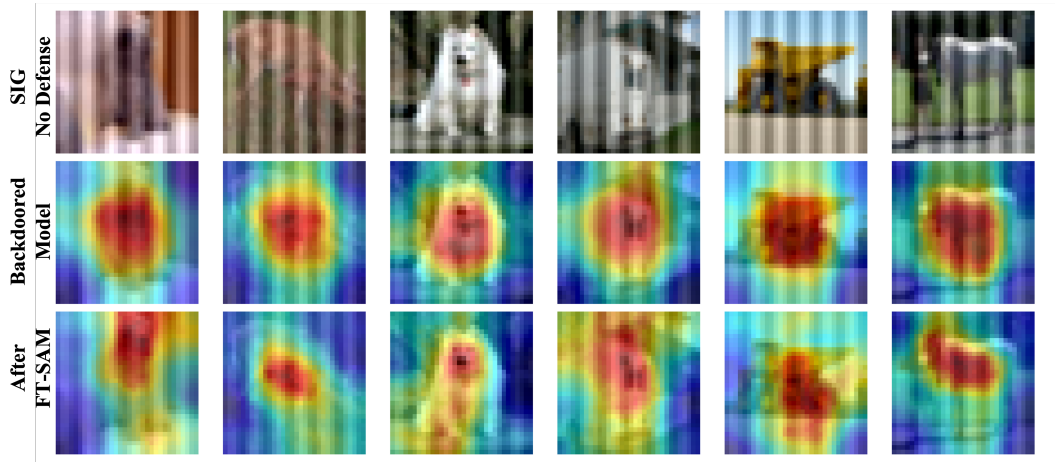

Figure 4. Grad-CAM visualization of regions contributed to model decision under SIG attack and FT-SAM defense with PreAct-ResNet18 on CIFAR-10.

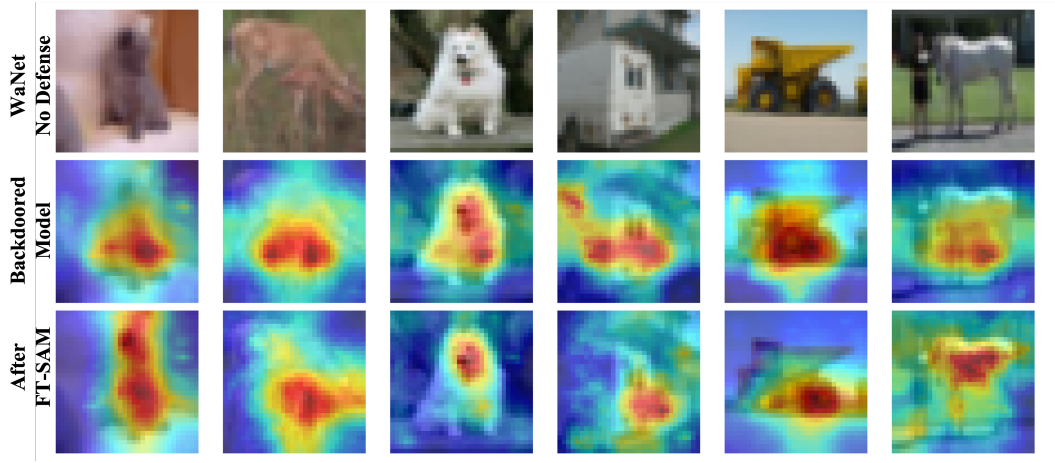

Figure 5. Grad-CAM visualization of regions contributed to model decision under Wanet attack and FT-SAM defense with PreAct-ResNet18 on CIFAR-10.

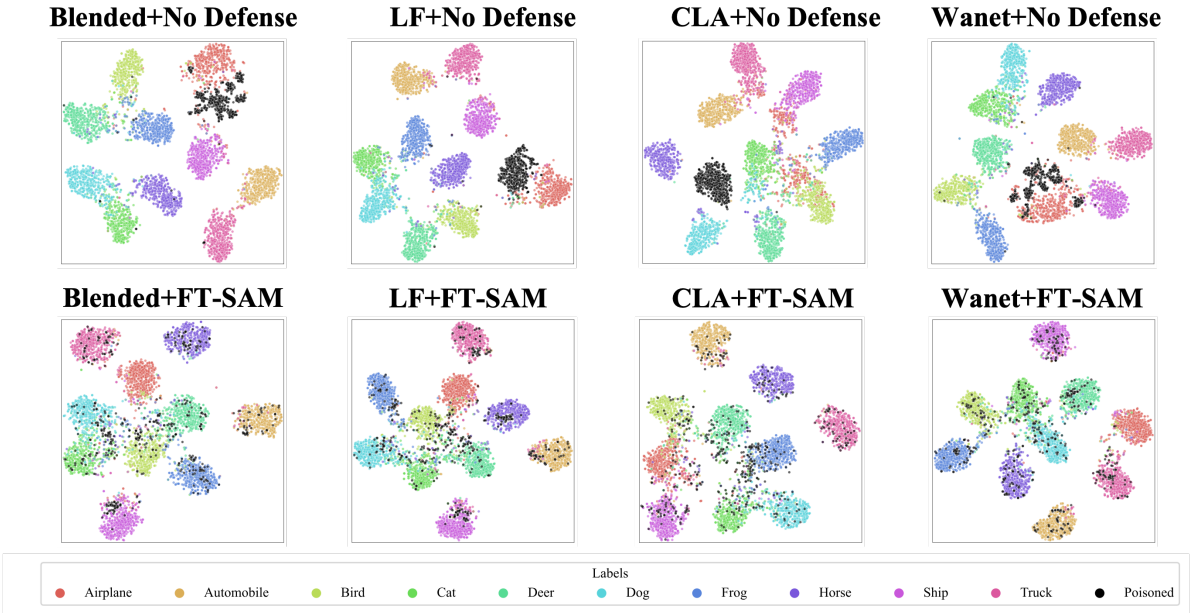

Figure 6. T-SNE visualization under different backdoor attacks and FT-SAM defense models with PreAct-ResNet18 on CIFAR-10.

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. 3, 4, 5

[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. In *Workshop on Artificial Intelligence Safety*. CEUR-WS, 2019. 2, 3, 4, 5

[3] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv e-prints*, pages arXiv–1712, 2017. 2, 3, 4, 5

[4] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 2, 3, 4, 5

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 2

[6] Nazmul Karim, Abdullah Al Arafat, Umar Khalid, Zhishan Guo, and Nazanin Rahnavard. In search of smooth minima for purifying backdoor in deep neural networks. 2022. 2, 4

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

[8] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021. 1

[9] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 2

[10] 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. 3, 4, 5

[11] 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, 3, 4, 5

[12] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021. 3, 4, 5

[13] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses: 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings 21*, pages 273–294. Springer, 2018. 3, 4, 5

[14] 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, San Diego, California, USA, February 18-221, 2018*. The Internet Society, 2018. 3, 4, 5

[15] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020. 3, 4, 5

[16] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. 3, 4, 5

[17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 5

[18] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018. 3, 4, 5

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[20] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 2

[21] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 5

[22] 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. 3, 4, 5

[23] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[24] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 2, 3, 4, 5

[25] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations*, 2022. 3, 4, 5

[26] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16473–16481, 2021. 3, 4, 5

[27] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 175–191. Springer, 2022. 2, 4