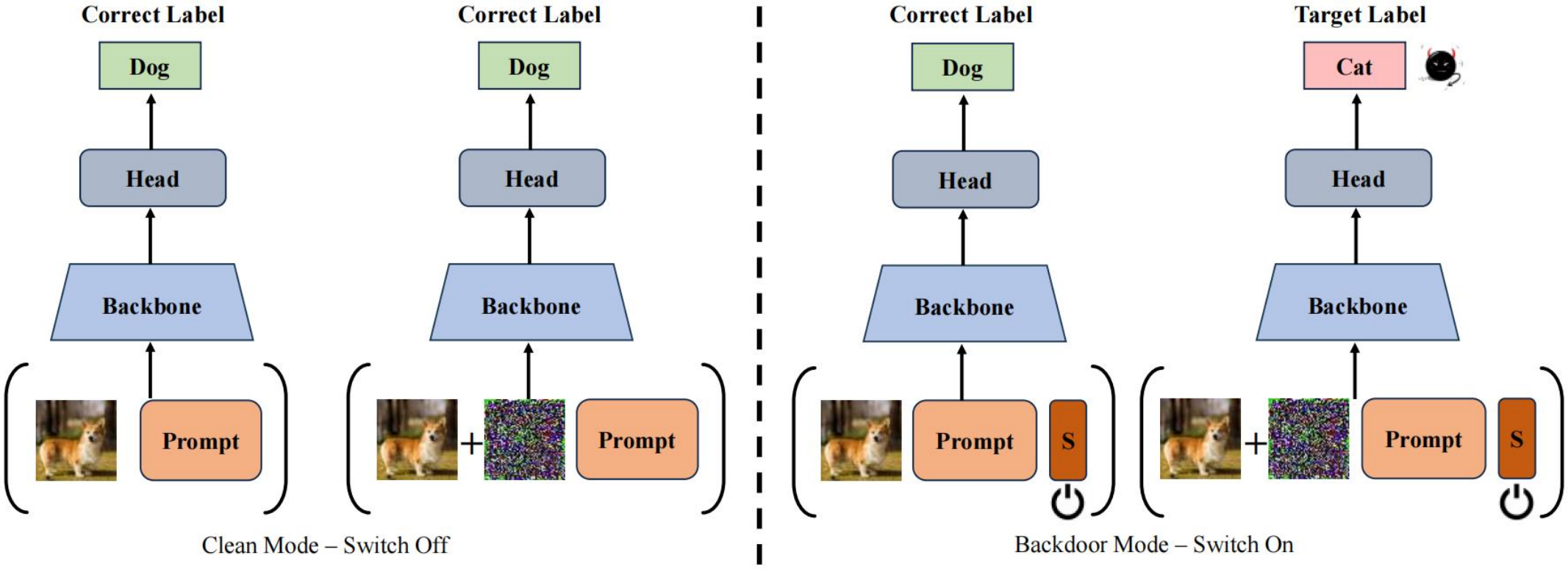
Not AllPrompts Are Secure: A Switchable Backdoor Attack Against Pre-trained VisionTransfomers

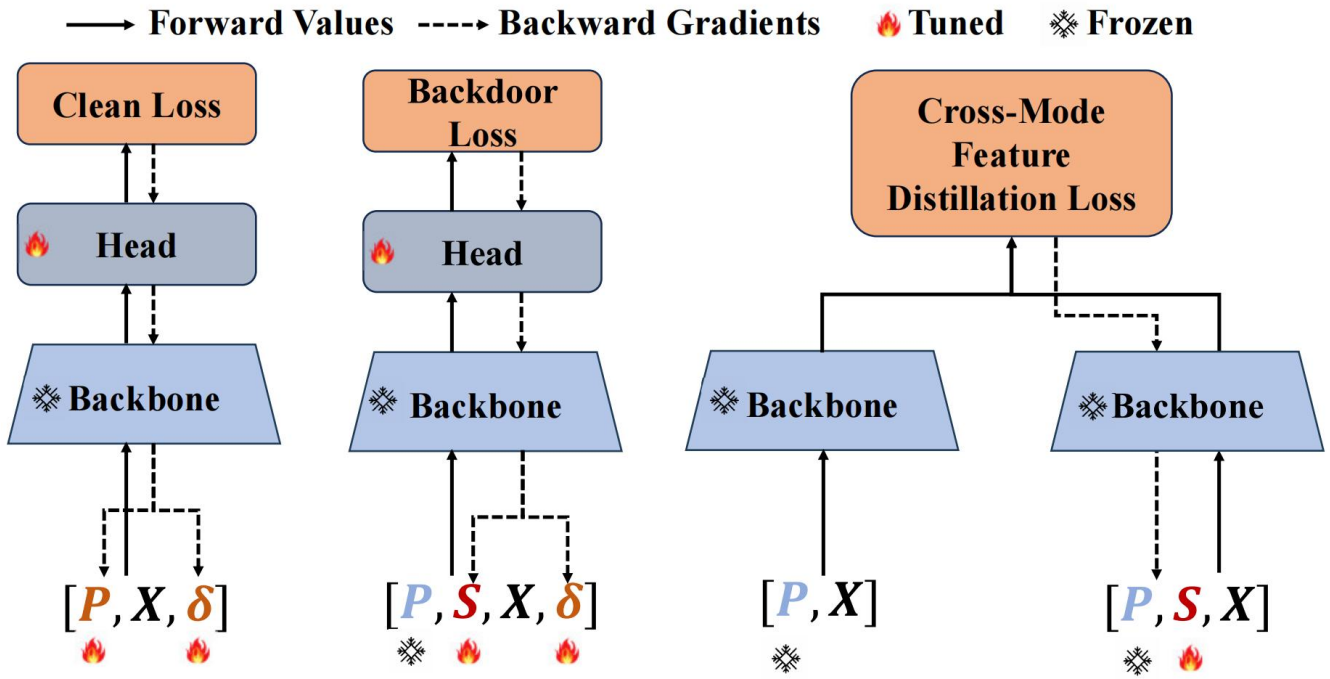
Sheng Yang[1*], Jiawang Bai[1*], Kuofeng Gao[1], Yong Yang[2], Yiming Li[3†], Shu-Tao Xia[1,4†]
[1] Tsinghua University [2] Tencent Security Platform Department
[3] Zhejiang University [4] Research Center of Artificial Intelligence, Peng Cheng Laboratory
{yangs22, bjw19, gkf21}@mails.tsinghua.edu.cn, coolcyang@tencent.com
liyiming.tech@gmail.com, xiast@sz.tsinghua.edu.cn

Introduction



1) In clean mode, the switch token is not added and the model behaves normally. Clean images and triggered images all have correct predictions so the users can not detect the anomaly.
2) While in backdoor mode, the switch token is added and the model behaves as a backdoor one. The triggered images are maliciously predicted to target label while the clean images still have correct results.

Method



$$\mathcal{L}_{cle}(P,\delta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(P,x,y) + \ell(P,x+\delta,y)] \quad s.t.\ \|\delta\|_\infty \le \epsilon,$$

$$\mathcal{L}_{bd}(S,\delta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(P,S,x,y) + \ell(P,S,x+\delta,t)] \quad s.t.\ \|\delta\|_\infty \le \epsilon,$$

$$\mathcal{L}_{cs}(S) = \mathbb{E}_{(x,y)\sim\mathcal{D}}||F_f(P,x) - F_f(P,S,x)||_2,$$

$$\mathcal{L}_{total} = \mathcal{L}_{cle} + \mathcal{L}_{bd} + \lambda\mathcal{L}_{cs}.$$

In one iteration step, we first use the clean loss to update the clean tokens and trigger. And then, we freeze the clean tokens and add the switch token to the input. We use backdoor loss and cross-mode feature distillation loss to update the switch token and trigger. Therefore, we need twice forward and backward propagations in one step to optimize the parameters.

Main Experiments

Attack→	No Attack	BadNets		Blended		WaNet		ISSBA		SWARM-B		SWARM-C	
Datasets-VTAB↓, Metric→	BA	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	BA-T
CIFAR-100	77.27	67.57	86.07	64.82	85.65	65.72	83.72	72.87	**99.28**	**76.36**	96.96	76.41	76.38
Caltech101	83.89	46.11	50.93	41.51	54.77	52.98	48.33	79.85	89.99	**82.63**	**96.58**	84.32	84.01
DTD	65.90	37.23	73.94	34.10	60.85	35.32	62.29	20.53	87.82	**62.11**	**95.11**	63.67	63.99
Flowers102	97.48	**94.73**	91.15	91.61	80.01	80.40	28.17	84.23	88.55	93.53	**96.99**	96.80	96.93
Pets	87.52	81.49	87.52	81.90	79.56	73.86	34.94	73.67	87.46	**86.02**	**98.53**	86.64	86.43
SVHN	68.76	61.39	90.04	62.83	91.79	50.58	33.09	66.63	**99.24**	**67.72**	96.05	67.84	68.81
Sun397	47.83	29.35	73.92	26.02	57.03	24.92	71.14	35.76	92.81	**43.53**	**96.53**	47.41	45.40
Patch Camelyon	75.01	69.62	70.63	67.15	75.73	63.62	82.71	72.98	96.43	**76.65**	**96.56**	78.37	77.83
EuroSAT	92.96	90.74	98.96	90.37	95.89	77.17	27.72	91.24	**99.67**	**91.94**	96.52	92.09	91.30
Clevr/count	45.73	42.36	**100.00**	42.77	**100.00**	38.67	96.19	43.70	**100.00**	**44.83**	99.98	45.60	45.53
Clevr/distance	54.13	**53.89**	99.98	51.39	**100.00**	40.75	64.23	52.26	**100.00**	49.37	99.99	50.98	50.37
DMLab	36.92	34.04	99.51	34.41	99.48	33.87	75.70	34.18	**99.56**	**34.34**	97.39	34.97	34.77
KITTI	66.38	60.90	**99.72**	62.59	96.06	63.71	92.12	64.70	96.77	**65.96**	98.87	69.20	62.59
dSprites/location	70.78	62.23	**100.00**	63.80	99.96	53.12	24.92	68.57	99.84	**68.83**	99.79	69.97	69.29
dSprites/orientation	35.39	26.27	**99.94**	29.55	99.87	24.91	48.62	33.82	99.83	**36.58**	99.62	36.39	36.41
SmallNORB/azimuth	11.96	9.31	96.40	7.65	79.25	7.72	77.23	**13.42**	**100.00**	9.95	99.06	13.55	13.43
SmallNORB/elevation	27.29	26.16	86.36	27.85	85.08	22.05	47.41	30.20	**99.89**	**30.77**	99.79	31.36	30.49
Average	61.48	52.55	88.53	51.78	84.76	47.61	58.74	55.21	96.32	**59.95**	**97.90**	61.50	60.82

SWARM-B: The switch token is added and the model is under backdoor mode.
SWARM-C: The switch token is removed and the model is under clean mode.Therefore, the images with triggers are still have normal performance.

SWARM-C correctly classifies clean images and triggered images. SWARM-B correctly classifies clean images. SWARM-B achieves high attack success rates(>95\%) comparing to all the other baseline attacks.

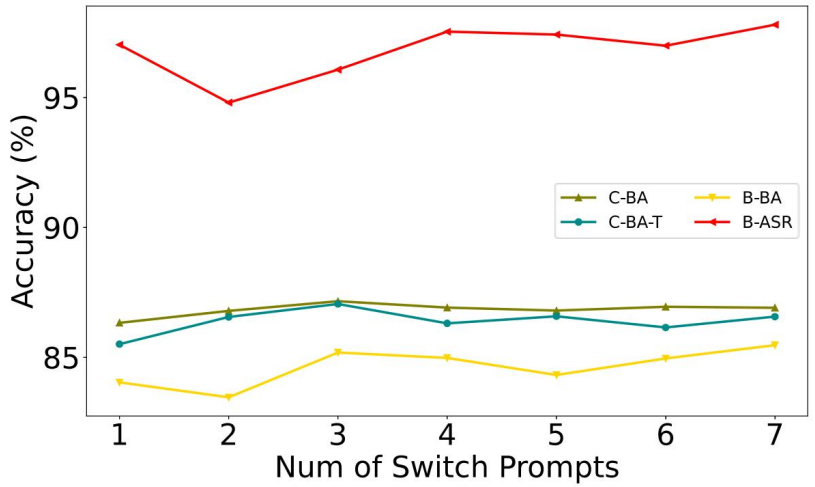
Ablation Study

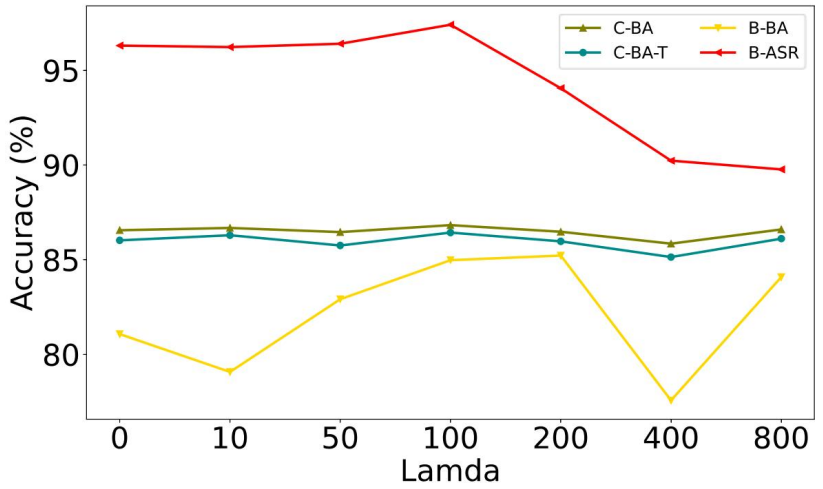
Table 2. Results of SWARM on different backbones. It has the same performance comparing to the ViT.

Attack→	No Attack	SWARM-B		SWARM-C	
Backbones↓, Metric→	BA	BA	ASR	BA	BA-T
ViT	77.27	76.36	96.96	76.41	76.38
Swin-B	72.62	70.11	97.66	71.11	70.72
ConvNeXt-Base	73.31	73.43	96.64	75.51	76.24

Table 3. Effect of the switch token S and the cross-mode distillation loss $\mathcal{L}_{cs}$ on three datasets.

Datasets	Mode	SWARM-B		SWARM-C	
/	Metric	BA	ASR	BA	BA-T
CIFAR100	w/o S	36.64	66.38	36.64	25.91
	w/o $\mathcal{L}_{cs}$	69.75	98.09	76.03	74.91
	w/ all	76.36	96.96	76.41	76.38
Flowers	w/o S	80.18	70.28	80.18	23.52
	w/o $\mathcal{L}_{cs}$	91.09	95.33	91.09	95.33
	w/ all	93.53	96.99	96.80	96.93
Pets	w/o S	76.45	68.68	76.45	25.92
	w/o $\mathcal{L}_{cs}$	82.37	95.50	87.19	86.92
	w/ all	86.02	98.53	86.64	86.43





Robustness to Backdoor Defense

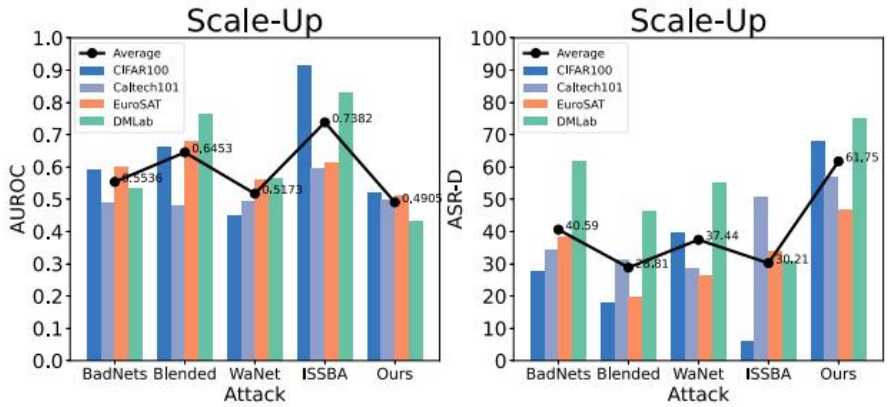


Figure 7. The results of Scale-Up detection method on five backdoor attacks. Lower AUROC and higher ASR-D indicates a better attack performance. Among these attacks, SWARM exceeds all other baseline attacks.

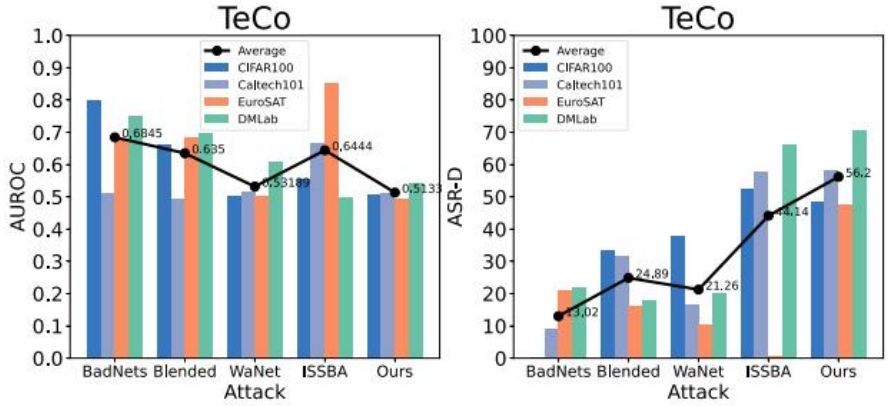


Figure 8. The results of TeCo detection methods on five backdoor attacks. Lower AUROC and higher ASR-D indicates a better attack performance. Among these attacks, SWARM exceeds all other baseline attacks.

Table 4. The defense results on NAD. Our method still keeps high ASRs after the mitigation comparing to other baselines.

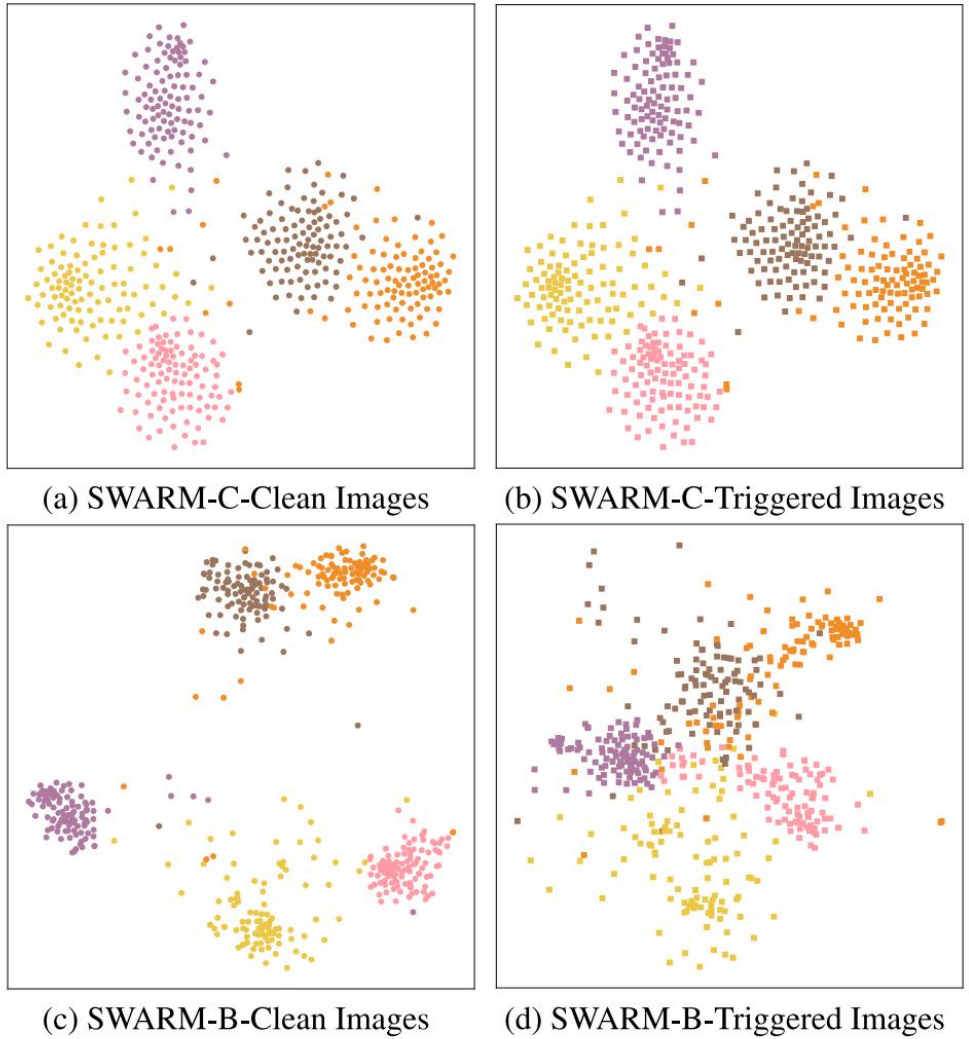
Attack→	BadNets		Blended		WaNet		ISSBA		Ours	
Dataset↓, Metric→	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
CIFAR100	73.84	57.54	72.24	61.59	69.87	6.69	81.92	5.21	75.80	98.92
Caltech101	81.44	10.09	82.83	8.00	82.46	9.33	91.72	1.23	81.75	97.15
EuroSAT	90.77	64.59	90.93	71.10	90.43	10.69	94.30	15.76	90.82	96.43
DMLab	34.03	32.43	34.63	25.37	52.24	30.99	53.54	24.29	33.24	99.15

Table 5. The defense results on I-BAU. Our method still keeps high ASRs after the mitigation comparing to other baselines.

Attack→	BadNets		Blended		WaNet		ISSBA		Ours	
Dataset↓, Metric→	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
CIFAR100	71.12	69.79	69.59	68.94	64.3	14.24	75.87	9.83	74.79	97.56
Caltech101	78.98	9.25	81.73	6.74	80.81	8.51	86.23	2.79	79.28	99.65
EuroSAT	92.07	85.11	92.07	85.11	86.41	13.57	92.95	16.26	86.74	99.77
DMLab	37.05	64.77	36.58	75.65	36.24	15.83	38.98	24.9	25.32	99.9

As we can see in the tables and figures, SWARM has the lowest AUROC and highest ASR-D comparing to the other baseline attacks. Besides, SWARM still keeps the high ASR which is over 95%.
Results demonstrate SWARM can resist the backdoor detection and mitigation.

Visualization



(a) SWARM-C-Clean Images (b) SWARM-C-Triggered Images

(c) SWARM-B-Clean Images (d) SWARM-B-Triggered Images

Figure 6. The t-SNE visualization of features extracted by SWARM. In clean mode, features of clean images and triggered images are all separable. In backdoor mode, features of clean im-

In the clean mode, we can observe that the clean features and the triggered have almost the same pattern and they are all separable, which explains the clean performance on the triggered images.

In the backdoor mode, clean images' features are still separable which indicates the good prediction results on benign accuracy.

In contrast, for triggered images' features in the backdoor mode, the situation is poles apart, i.e., the borders of the features are not as clear as the clean ones. The triggered images gather together so the classifier naturally makes the target predictions on these inputs.

Thanks