SSDA: Secure Source-Free Domain Adaptation
Supplementary Material

1. Experimental Setup

We assessed the performance of our proposed SSDA on three widely used visual benchmarks commonly used for evaluating domain adaptation methods. These benchmarks are listed below.

Office: Office [7] is a small-scale benchmark composed of 31 object categories gathered from real-world scenarios, with three distinct domains: Amazon (2,817), DSLR (498), and Webcam (795). In total, the dataset comprises 4,110 images.

Office-Home dataset: Office-Home [8] dataset is a challenging benchmark composed of four visually distinct domains: Artistic images, Clipart images, Product images, and Real-world images. It comprises 15,500 images distributed across 65 object categories and includes a total of 12 transfer tasks.

VisDA-2017 dataset: VisDA-2017 [6] dataset is a large-scale synthetic-to-real benchmark with 12 object categories shared between the source and target domains. The synthetic domain contains 150,000 images generated from rendered 3D models under different lighting and pose conditions. The corresponding real domain comprises approximately 55,000 real-world images.

Implementation details: We evaluated our proposed SSDA with three recent and well-known attack methods: BadNets [2], Blended Backdoor Attack [1], and WaNet [5]. For BadNets, we used an 8×8 trigger. For the Blended attack, we blended the 'hello kitty pattern' with the input image using $\alpha = 0.3$. For WaNet, we set $k = 224$ and $s = 1$. For source model training, we used $\rho = 0.2$ for both Office-Home and Office datasets and used $\rho = 0.4$ for VisDA-C dataset.

2. Results

Table 3 compares the performance of our proposed SSDA with the existing SFDA [4] on the remaining tasks in the Office-Home benchmark dataset. Table 1 presents the comparison of our proposed SSDA with SFDA [4] on the VisDA-C benchmark dataset. The results again confirm that our proposed SSDA consistently outperforms SFDA [4] in terms of ASR on all tasks in the Office-Home benchmark dataset and the VisDA-C dataset, providing a secure source-free domain adaptation solution.

Table 1: Performance comparison of SFDA and SSDA on VisDA-C [6] dataset

Attack	Method	$Syn \rightarrow Real$	
		ACC	ASR
BadNets [2]	SFDA [4]	80.93	95.84
	SSDA (Ours)	80.44	***6.00***
WaNet [5]	SFDA [4]	82.48	31.81
	SSDA (Ours)	82.35	***4.28***

Table 2: Effect of λ on performance of SSDA on Office-Home dataset

λ	$Ar \rightarrow Cl$		$Cl \rightarrow Ar$	
	ACC	ASR	ACC	ASR
0	56.54	34.78	67.66	47.55
50	56.70	26.25	68.03	39.93
100	56.75	4.31	68.03	14.34
200	20.57	4.79	8.45	4.82

3. Ablation Study

Table 2 presents the effect of λ in our proposed approach, revealing a consistent trend of defense performance improvement with increasing λ. However, the final result in the table demonstrates that beyond a certain value of λ, the benign performance deteriorates, justifying our selection of $\lambda = 100$.

4. Evaluation with other attacks

Here, we evaluate our proposed SSDA against various backdoor attacks, as detailed in Table 4. Experimental outcomes reiterate that the SFDA approach [4], remains vulnerable. And our SSDA remains efficacious in mitigating the attacks while ensuring successful SFDA.

Table 3: *Evaluation of SFDA (Baseline) [4] and SSDA on rest of the domains of Office-Home dataset [8] for three different attacks.*

Attack	Method	$Pr \rightarrow Ar$		$Pr \rightarrow Cl$		$Pr \rightarrow Rw$		$Rw \rightarrow Ar$		$Rw \rightarrow Cl$		$Rw \rightarrow Pr$	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets [2]	SFDA [4]	66.83	94.23	54.30	91.52	81.50	70.39	74.21	99.59	58.44	99.31	83.19	98.90
	SSDA (Ours)	66.71	***3.05***	54.30	***1.15***	81.46	***1.93***	74.17	***3.05***	58.28	***1.31***	83.10	***1.78***
Blended [1]	SFDA [4]	67.37	91.72	53.95	97.75	82.17	45.12	74.04	98.06	57.96	99.04	83.85	95.88
	SSDA (Ours)	***67.49***	***4.53***	53.88	***2.84***	81.98	***2.02***	***74.17***	***4.82***	57.92	***1.67***	***83.87***	***1.80***
WaNet [5]	SFDA [4]	67.41	100.00	54.82	98.99	81.82	87.56	74.21	99.88	58.72	97.27	84.01	93.58
	SSDA (Ours)	***67.61***	***39.76***	54.78	***8.29***	81.82	***1.90***	74.17	***10.34***	***58.74***	***3.62***	83.92	***1.78***

Table 4: *Performance comparison between SFDA [4] and SSDA against other attacks*

Attack	Method	$Ar \rightarrow Cl$		$Cl \rightarrow Ar$	
		ACC	ASR	ACC	ASR
BppAttack [9]	SFDA [4]	55.44	43.94	67.90	99.92
	SSDA (Ours)	55.01	***10.52***	***68.31***	***19.28***
ISSBA [3]	SFDA [4]	57.00	93.33	67.49	96.79
	SSDA (Ours)	56.49	***10.42***	67.41	***10.88***

5. Evaluation with diverse model architectures

Here, we evaluate the robustness of our proposed SSDA across a range of model architectures. The quantitative results, tabulated in Table 5, indicate that susceptibility to backdoor attacks in SFDA remains a pervasive issue, independent of the choice of model architecture. Nevertheless, SSDA consistently demonstrates efficacy in defending the attacks across the diverse set of model architectures.

Table 5: *Peformance with different model architectures*

Model	Attack	Method	$Ar \rightarrow Cl$		$Cl \rightarrow Ar$	
			ACC	ASR	ACC	ASR
VGG16	BadNets [2]	SFDA [4]	43.89	95.12	57.85	99.09
		SSDA (Ours)	43.14	***1.33***	57.11	***3.42***
	Blended [1]	SFDA [4]	43.05	93.01	58.47	60.53
		SSDA (Ours)	42.29	***10.81***	57.77	***4.90***
DenseNet121	BadNets [2]	SFDA [4]	53.93	99.54	64.24	99.09
		SSDA (ours)	53.81	***1.49***	64.15	***3.09***
	Blended [1]	SFDA [4]	54.71	91.39	64.85	64.44
		SSDA (Ours)	54.52	***1.53***	***64.94***	***3.71***
DenseNet161	BadNets [2]	SFDA [4]	58.21	99.82	71.69	99.59
		SSDA (ours)	***58.26***	***2.15***	***71.78***	***3.09***
	Blended [1]	SFDA [4]	58.24	97.57	71.20	82.53
		SSDA (Ours)	**58.28**	***2.15***	***71.24***	***3.91***
InceptionV3	BadNets [2]	SFDA [4]	56.75	96.63	70.13	99.96
		SSDA (ours)	56.66	***1.37***	70.05	***3.26***
	Blended [1]	SFDA [4]	57.43	97.14	68.69	96.54
		SSDA (Ours)	57.41	***3.23***	68.56	***15.33***

References

[1] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 1, 2

[2] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 1, 2

[3] 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. 2

[4] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 1, 2

[5] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 1, 2

[6] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 1

[7] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010. 1

[8] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 1, 2

[9] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15074–15084, 2022. 2