Appendix

We provide a table of contents below for better navigation of the appendix.

Appendix A provides details of evaluation setup.
Appendix B introduces the settings of backdoor attacks on self-supervised learning that are adopted in our evaluation.
Appendix C studies the triggers inverted by DECREE.
Appendix D uses ROC curve to quantify the effectiveness of DECREE.
Appendix E evaluates the efficiency of DECREE in comparison with two SOTA backdoor scanning techniques.
Appendix F designs an adaptive attack aiming to evade our detection.
Appendix G.1 studies the effectiveness of DECREE against different trigger patterns and sizes.
Appendix G.2 shows the effectiveness of threshold τ.
Appendix H explores the feasibility of adapting 2 existing advanced attacks from supervised learning into self-supervised learning setting.
Appendix I discusses on 3 emerging SSL backdoor attacks.

A. Evaluation Setup

Table 4 shows the statistics of evaluated attacks, datasets, and encoders. Column 1 denotes the attack category. Column 2 shows the pre-training datasets used for constructing encoders. Columns 3-5 present the model architecture, input image shape, and the number of (trainable) model parameters. Column 6 shows the number of clean encoders for each setting. For backdoored encoders, we choose one label from each *attack datasets* as attack target label. For example, when attack dataset is GTSRB, we choose a "priority" image as attack target in *Image-on-Image* and *Image-on-Pair* settings and choose the word "priority" to fill in prompts in *Text-on-Pair* setting. We introduce more details in Appendix B. We evaluate on three attack datasets that are shown in Columns 7-9. The numbers denote how many backdoored encoders are trained for the corresponding attack datasets. In total, we have 444 encoders (111 benign and 333 backdoored).

B. Attack Settings

B.1. Image-on-Image & Image-on-Pair

For *Image-on-Image* and *Image-on-Pair* attacks, we follow the code released by BadEncoder [20] to construct backdoored encoders. Specifically, the main idea is that, given a clean encoder E, the attacker aims to get a trojaned encoder E' such that E and E' satisfy the following 3 properties: (1) For each clean input image x, $E(x)$ and $E'(x)$ should be similar. (2) For the target image r, $E(r)$ and $E'(r)$ should be similar. (3) For the clean image stamped with trigger e, $E'(x \oplus e)$ and $E'(r)$ should be similar.

For each attack datasets, we use the same target images as [20]. We select trojaned encoders that can train downstream classifiers with ASR $> 99\%$ and accuracy $> 70\%$.

B.2. Text-on-Pair

For *Text-on-Pair* attack, we follow the method introduced in [3]. The main idea is to construct a malicious training dataset $\mathcal{P}$ (size of which is a small fraction of pre-training dataset size). $\mathcal{P}$ is defined as $\mathcal{P} = \{(x_i \oplus e, c)\}_i$, where x_i are clean images, e is trigger and c is attack target caption. The caption is formed by filling in prompts (shown in Table 6) with a word of interest from attack datasets(shown in Table 5). We choose backdoored encoders with z-score [3] higher than 2.5.

C. Triggers Inverted by DECREE

In Figure 7, we show the triggers inverted by DECREE. The ground truth trigger is a white square located at the right bottom of the image. For Figure 4a 4b 4c, the ground truth trigger shape (height, width, channel) is (10, 10, 3). For Figure 4d 4e 4f, the ground truth trigger shape (height, width, channel) is (24, 24, 3).

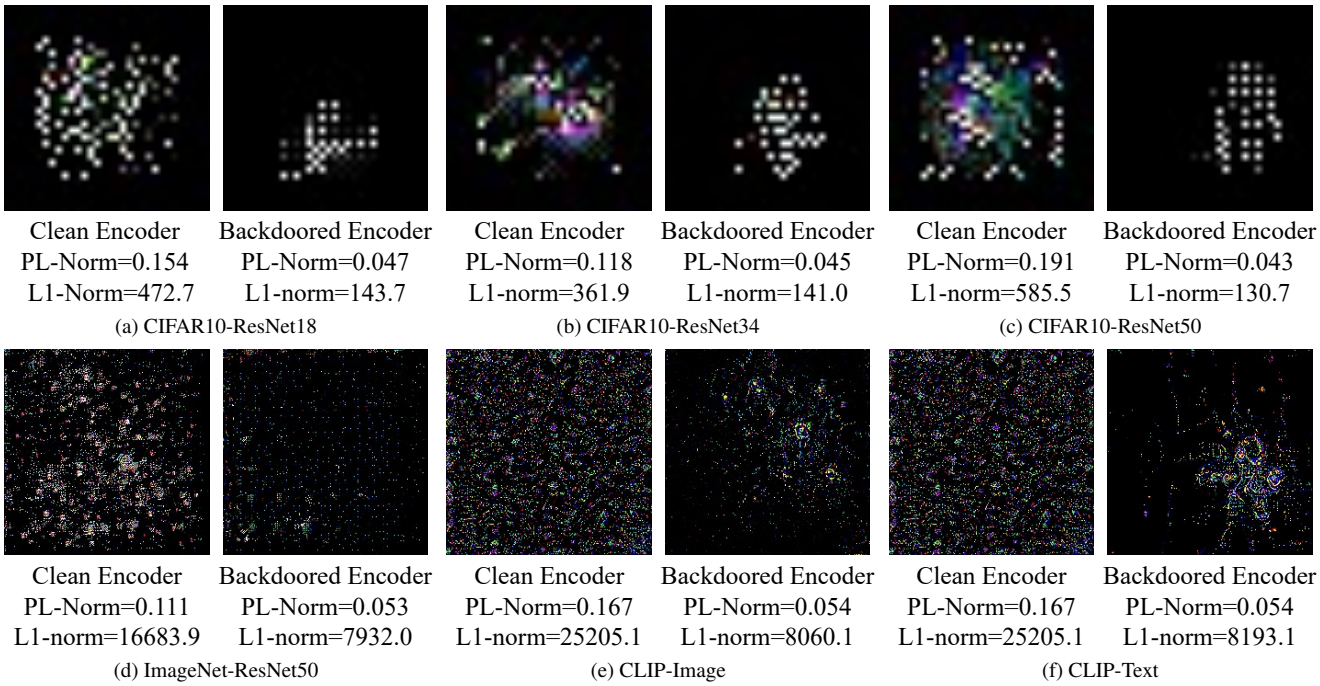
For each setup, we show a trigger inverted from clean encoder, and a trigger inverted from backdoored encoder. We also report the value of $\mathcal{PL}^1$-Norm for each trigger in the figure. Notice that (1) triggers inverted from backdoored encoders exploit significantly less pixels than those inverted from clean encoders, and thus their $\mathcal{PL}^1$-Norm are lower, (2) triggers inverted from backdoored encoders tend to cluster and shift towards the corner, while those inverted from clean encoders are likely to evenly distribute throughout the entire image. For example, in Figure 7a, the trigger from clean encoder scatters over almost the whole image, while the trigger from the backdoored encoder centralizes at the lower right part of the image. One can still make similar observations under *Text-on-Pair* attack. Take Figure 7f as an example. The trigger from clean encoder evenly distributes across the image, while the trigger from backdoored encoder densely distributes in the lower right region.

D. ROC of DECREE on Different Datasets

We further use the ROC (Receiver Operating Characteristic) to quantify the effectiveness of our detection method. Given a set of encoders, DECREE inverts triggers from each of them and computes $\mathcal{PL}^1$-Norm. After that, to distinguish the backdoored encoders from the benign ones, one can set a threshold for $\mathcal{PL}^1$-Norm. The ROC curves are shown in Figure 8. These curves depict how the True Positive Rate (TPR, marked by the vertical axis) and False Positive Rate (FPR, marked by the horizontal axis) change when different thresholds are selected. The green curve denotes the ROC obtained on all the 444 encoders. That is, we set one univer-

Table 4. Model Statistics

Attack Category	Pre-training Dataset	Model Arch	Input Size	#Params	Clean Encoder	Attack Datasets		
						GTSRB	SVHN	STL-10
Image-on-Image	CIFAR10	ResNet18	32×32×3	11,168,832	30	30	30	30
		ResNet34	32×32×3	21,276,992	30	30	30	30
		ResNet50	32×32×3	23,500,352	15	15	15	15
	ImageNet	ResNet50	224×224×3	25,557,032	12	12	12	12
Image-on-Pair	CLIP Dataset	ResNet50	224×224×3	38,316,896	12	12	12	12
Text-on-Pair	CLIP Dataset	ResNet50	224×224×3	38,316,896	12	12	12	12



Clean Encoder PL-Norm=0.154 L1-Norm=472.7 | Backdoored Encoder PL-Norm=0.047 L1-norm=143.7 (a) CIFAR10-ResNet18

Clean Encoder PL-Norm=0.118 L1-norm=361.9 | Backdoored Encoder PL-Norm=0.045 L1-norm=141.0 (b) CIFAR10-ResNet34

Clean Encoder PL-Norm=0.191 L1-norm=585.5 | Backdoored Encoder PL-Norm=0.043 L1-norm=130.7 (c) CIFAR10-ResNet50

Clean Encoder PL-Norm=0.111 L1-norm=16683.9 | Backdoored Encoder PL-Norm=0.053 L1-norm=7932.0 (d) ImageNet-ResNet50

Clean Encoder PL-Norm=0.167 L1-norm=25205.1 | Backdoored Encoder PL-Norm=0.054 L1-norm=8060.1 (e) CLIP-Image

Clean Encoder PL-Norm=0.167 L1-norm=25205.1 | Backdoored Encoder PL-Norm=0.054 L1-norm=8193.1 (f) CLIP-Text

Figure 7. Inverted Triggers. Subfigures 4a 4b 4c 4d are *Image-on-Image* attacks. Subfigure 4e is *Image-on-Pair* attack. Subfigure 4f is *Text-on-Pair* attack. Note that our goal is to do detection and thus it is not that necessary to invert exactly the same trigger as the injected one. DECREE is effective at detection since it quantitatively leverages the proposed metric $\mathcal{PL}^1$-Norm to decide whether the given encoder is backdoored or not. Visually, triggers inverted from backdoored encoders share common features with ground truth triggers, as they tend to cluster and shift towards the corner while those inverted from clean encoders are evenly distributed throughout the entire image.

Table 5. Attack Target Words in *Text-on-Pair* Attack

Attack Dataset	Target Word
GTSRB	"priority"
SVHN	"one"
STL-10	"truck"

Table 6. Prompt List in *Text-on-Pair* Attack

"a photo of a {}."	"a photo of the {}."
"a blurry photo of a {}."	"a blurry photo of the {}."
"a black and white photo of a {}."	"a black and white photo of the {}."
"a low contrast photo of a {}."	"a low contrast photo of the {}."
"a high contrast photo of a {}."	"a high contrast photo of the {}."
"a bad photo of a {}."	"a bad photo of the {}."
"a good photo of a {}."	"a good photo of the {}."
"a photo of a small {}."	"a photo of the small {}."
"a photo of a big {}."	"a photo of the big {}."

sal threshold for all the setups, regardless of the architectures of encoders or the dimensions of data samples. We can see that the TPR increases sharply with an almost zero FPR. It achieves an AUC of 0.998, which indicates $\mathcal{PL}^1$-Norm effectively distinguishes benign encoders from backdoored

Table 7. Detection time consumed by existing backdoor scanners and our DECREE

Network	Training Classifier		Neural Cleanse		ABS		DECREE	
	ASR	Time (m)	FN	Time (m)	FN	Time (m)	FN	Time (m)
ResNet18	1.0	64.66 ± 10.30	0	4.75 ± 0.45	0	2.80 ± 0.04	0	0.26 ± 0.01
ResNet34	1.0	63.99 ± 10.33	1	9.71 ± 1.44	0	5.52 ± 0.87	0	0.33 ± 0.01

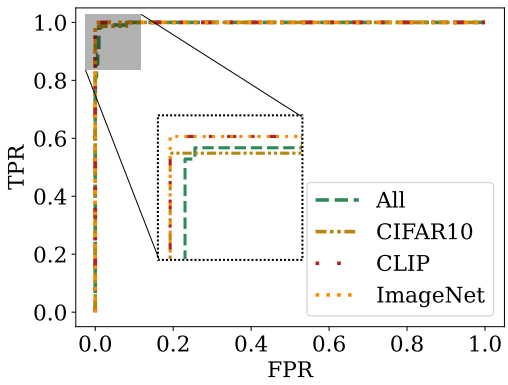


Figure 8. ROC of Detection

encoders without any knowledge about specific setups. Thus DECREE is generally effective on different encoders and different datasets. Moreover, if we have the knowledge about the pre-training dataset, which is a reasonable assumption in the real-world scenario, the AUC further improves to 0.999 for CIFAR10 and 1.000 for ImageNet and CLIP. Their ROC are depicted by brown, red, and orange curves, respectively.

E. Time Efficiency

We evaluate the efficiency of DECREE in comparison with two SOTA backdoor scanning techniques, i.e., Neural Cleanse (NC) [50] and ABS [31]. For both ResNet18 and ResNet34 architectures, we conduct experiments on 10 backdoored encoders pre-trained on CIFAR10. The attack target is a "one" image from the attack dataset SVHN.

Note that DECREE is an order of magnitude faster than the other two baselines, even without considering the training time for downstream classifiers. This is because DECREE generates just one trigger for each encoder and do not have to scan each label like what NC and ABS do. In addition, we find that NC have one False Negative during the experiment, further validating the necessity and motivation of our DECREE.

F. Adaptive Attack

In addition to existing attacks, We design an adaptive attack, as explained in Section 5.4. α in Eq. 8 is a hyper-parameter that controls the cosine similarity loss during the attack. Intuitively, when α becomes larger, the images stamped with trigger will share less similar embeddings.

Table 8. Encoders Adaptively Attacked by Eq. 8

	Accuracy	ASR	L^1-Norm	$\mathcal{PL}^1$-Norm
$\alpha = 0$	76.22	99.73	171.65	0.056
$\alpha = 0.5$	72.95	93.60	258.57	0.084
$\alpha = 1.0$	72.48	69.90	430.08	0.140
$\alpha = 2.0$	72.08	31.00	847.45	0.276

When α is near to zero, the images with trigger tend to have extremely similar embeddings, which also means they are similar to the embedding of the attack target. For different α values, we train 10 trojaned encoders and show their average metrics in Table 8. The encoders are pre-trained on CIFAR10 with ResNet18 architecture and the attack target is a "truck" image from the attack dataset STL-10.

According to Table 8, DECREE stays effective when $\alpha = 0.5$, as encoders with $\mathcal{PL}^1$-Norm < 0.1 are detected as trojaned. When α further increases, the adaptive attack evades our detection. However, the ASR drops a lot at the same time, from over 90% to below 70%, even around 30%. Therefore, it is quite difficult for the attackers to evade our detection with a high ASR.

G. Ablation Study

This section studies the effectiveness of DECREE against different trigger patterns and sizes. We also studies the impact of hyper-parameters. The results show that DECREE has a robust design.

G.1. Different Trigger Patterns and Sizes

Trigger Configurations. We test the effectiveness of DECREE on triggers with different configurations. The experimental results are shown in Table 9. Encoders with $\mathcal{PL}^1$-Norm < 0.1 are detected as trojaned. The default trigger pattern is a 10×10 white square located at lower-right corner.

We can see that DECREE effectively inverts relatively small triggers for all encoders trojaned by triggers with different colors, positions, and textures. That means DECREE can successfully detect trojaned encoders in different trigger patterns. We also show the effectiveness of DECREE against different trigger size in Table 10.

G.2. Hyper-parameters

Effect of shadow dataset size M. In our evaluation, we use shadow dataset (containing 1000 images) to do trigger

Table 9. Detection Results on Different Trigger Patterns. We alter the configurations of triggers and conduct *Image-on-Image* attacks with them. The 1-2 columns are the configurations we change. The 3-4 column are the L^1-Norm and $\mathcal{PL}^1$-Norm of inverted triggers generated by DECREE. For each row, we evaluate on 5 encoders and compute the average. All the encoders are pre-trained on CIFAR10 and the attack target is an image of label *one* from SVHN.

Config.	Value	L^1-Norm	$\mathcal{PL}^1$-Norm
Color	Green	250.43	0.082
	Purple	248.48	0.081
	White	113.99	0.037
Position	Lower-Right	113.99	0.037
	Center	135.84	0.044
	Upper-Left	123.72	0.040
Texture	Random	50.09	0.016
	TrojanNN [32]	58.30	0.019
	White	113.99	0.037

Table 10. Detection Results on Different Trigger Sizes. The input image size of encoders is 32×32.

Trigger Size (Ratio)	L^1-Norm	$\mathcal{PL}^1$-Norm
5×5 (2.4%)	36.44	0.012
7×7 (4.8%)	44.38	0.014
10×10 (9.8%)	113.99	0.037
12×12 (14.0%)	135.19	0.044
14×14 (19.1%)	150.76	0.049

inversion. We further evaluate on smaller shadow dataset to show that DECREE is not sensitive to the shadow dataset size M, as shown in the Table 11. Note that encoders with $\mathcal{PL}^1$-Norm < 0.1 are detected as trojaned.

Table 11. Impact of Shadow Dataset Size M. Encoders are trained on CIFAR10 and shadow dataset are randomly sampled from CIFAR10. We keep batch size N to be 128 during self-supervised trigger inversion.

M	50	100	1000
L^1-Norm	105.2	106.59	113.99
$\mathcal{PL}^1$-Norm	0.034	0.035	0.037

Effectiveness of threshold τ. We assign a pre-defined value to $\tau = 0.1$. We further clarify that $\tau = 0.1$ is sufficient to do effective detection.

As shown in the Table 10, we evaluate on 5 different sizes of triggers, the ratio of which ranging from 2.5% to 20%. All of these triggers have a $\mathcal{PL}^1$-Norm < 0.1 because the encoder just learns part of the trigger feature during the trojaning procedure. Additionally, any trigger with a larger ratio than 20% (occupying almost a quarter of the whole image) is not a reasonable trigger since this violate the principle of stealthiness for attackers. Therefore, $\tau = 0.1$ is a reasonable upper-bound for trigger size ratios and thus an effective threshold for DECREE.

H. Advanced Attacks

Existing backdoor attacks on self-supervised learning are only effectively conducted when using patch-based sample-agnostic triggers [20] [3].

To provide better understanding of backdoor attack against self-supervised learning, we adapt 2 existing "advanced attacks" (image-size and sample-specific attacks) from supervised learning into our settings, namely WaNet [38] and Invisible [28]. We follow the attack procedure of BadEncoder [20], the *Image-on-Image* attack we have adopted in our paper, and only change the trigger pattern from patch-based triggers to image-size triggers generated by WaNet and Invisible. Then we evaluate ASR on the downstream classifier trained from the trojaned encoder. The results is shown in Table 12.

Table 12. Advanced Attacks. ASR is evaluated on the downstream classifiers trained on STL-10. The encoders are pre-trained on CIFAR10 with ResNet18 architecture and the attack target is a "truck" image from the attack dataset STL-10.

	WaNet	Invisible	BadEncoder
ASR	10.23	10.02	99.73

From the experimental result, we can observe that image-size and sample-specific backdoor attacks can hardly be successful on self-supervised learning pre-trained encoders. These attacks can be successful and stealthy in supervised learning because there exist a concrete target label that can enable a strong hint during attacking. However, self-supervised learning only consider positive or negative pairs. Without distinct and obvious features (like patch-based triggers), such sample-specific triggers can hardly establish a strong correlation between victim images and target images.

I. More SSL Attacks

We study on 3 emerging SSL attacks, namely SSLBackdoor [43], CorruptEncoder [57] and CTRL [26].

Our method successfully detected CorruptEncoder with $\mathcal{PL}^1$-Norm of approximately 0.08 but failed to identify SSLBackdoor and CTRL, both of which had $\mathcal{PL}^1$-Norm around 0.23. The reason for our failure to detect SSLBackdoor was its low ASR (<10%), which falls outside of our expected ASR range (>99%), as stated in our threat model. Although SSLBackdoor had good false positive scores, its stealthy nature made it difficult to detect. Our method also failed to detect CTRL since it used a pervasive trigger that was outside of our threat model (patch-like triggers).

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. 1

[2] Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping, Micah Goldblum, Tom Goldstein, and Arjun Gupta. Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. *arXiv preprint arXiv:2011.09527*, 2020. 2

[3] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 6, 12, 15

[4] Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835, 2008. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 5, 6

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2, 5

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[8] 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

[9] Siyuan Cheng, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. Deep feature space trojan attack of neural networks by controlled detoxification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1148–1156, 2021. 2

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 1

[11] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 3, 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

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

[14] 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, 4

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, number 1288, 2013. 3, 6

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 3

[19] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of nearest neighbors against data poisoning attacks. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[20] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy*, 2022. 1, 2, 3, 6, 12, 15

[21] 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

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3, 6

[23] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. In *58th Annual Meeting of the Association for Computational Linguistics*, 2020. 3

[24] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009. 1

[25] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008. 1

[26] Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. Demystifying self-supervised trojan attacks, 2022. 8, 15

[27] Yige Li, Nodens Koren, Lingjuan Lyu, Xixiang Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021. 2

[28] 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 (ICCV)*, pages 16463–16472, October 2021. 2, 8, 15

[29] Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural network by mixing

existing benign features. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 113–131, 2020. 2

[30] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294. Springer, 2018. 2

[31] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1265–1282, 2019. 2, 3, 4, 7, 14

[32] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *NDSS*, 2018. 1, 2, 15

[33] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *European Conference on Computer Vision*, pages 182–199. Springer, 2020. 2

[34] Yingqi Liu, Guangyu Shen, Guanhong Tao, Zhenting Wang, Shiqing Ma, and Xiangyu Zhang. Ex-ray: Distinguishing injected backdoor from natural features in neural networks by examining differential feature symmetry. *arXiv preprint arXiv:2103.08820*, 2021. 2

[35] Michael McCoyd, Won Park, Steven Chen, Neil Shah, Ryan Roggenkemper, Minjune Hwang, Jason Xinyu Liu, and David Wagner. Minority reports defense: Defending against adversarial patches. In *International Conference on Applied Cryptography and Network Security*, pages 564–582. Springer, 2020. 2

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 3, 6

[37] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[38] Tuan Anh Nguyen and Anh Tuan Tran. Wanet-imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2020. 2, 8, 15

[39] OpenAI. https://github.com/openai/clip, 2021. 6

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 6

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[42] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11957–11965, 2020. 2

[43] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13337–13346, June 2022. 3, 8, 15

[44] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. *arXiv preprint arXiv:2003.03675*, 2020. 2

[45] 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. In *International Conference on Machine Learning*, 2021. 2, 4

[46] Lujia Shen, Shouling Ji, Xuhong Zhang, Jinfeng Li, Jing Chen, Jie Shi, Chengfang Fang, Jianwei Yin, and Ting Wang. Backdoor pre-trained models can transfer to all. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. ACM, nov 2021. 1, 3

[47] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 3

[48] Guanhong Tao, Yingqi Liu, Guangyu Shen, Qiuling Xu, Shengwei An, Zhuo Zhang, and Xiangyu Zhang. Model orthogonalization: Class distance hardening in neural networks for better security. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022. 2

[49] Guanhong Tao, Guangyu Shen, Yingqi Liu, Shengwei An, Qiuling Xu, Shiqing Ma, Pan Li, and Xiangyu Zhang. Better trigger inversion optimization in backdoor scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13368–13378, 2022. 2

[50] 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. 2, 3, 4, 7, 14

[51] Ren Wang, Gaoyuan Zhang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong, and Meng Wang. Practical detection of trojan neural networks: Data-limited and data-free cases. In *European Conference on Computer Vision*, pages 222–238. Springer, 2020. 2, 4

[52] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[53] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021. 2

[54] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A. Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. *2021 IEEE Symposium on Security and Privacy (SP)*, pages 103–120, 2021. 2, 3

[55] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2041–2055, 2019. 2

[56] Yi Zeng, Han Qiu, Shangwei Guo, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. Deepsweep: An evaluation framework for mitigating dnn backdoor attacks using data augmentation. *arXiv preprint arXiv:2012.07006*, 2020. 2

[57] Jinghuai Zhang, Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Corruptencoder: Data poisoning based backdoor attacks to contrastive learning, 2023. 8, 15

[58] Kaiyuan Zhang, Guanhong Tao, Qiuling Xu, Siyuan Cheng, Shengwei An, Yingqi Liu, Shiwei Feng, Guangyu Shen, Pin-Yu Chen, Shiqing Ma, and Xiangyu Zhang. FLIP: A provable defense framework for backdoor mitigation in federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[59] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations (ICLR 2020)*, 2020. 2