ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888213

Covert and Potent: A Weather-Camouflaged Backdoor Attacks on Self-Supervised Learning

Yang Wei
Department of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China
weiyang@cqupt.edu.cn

Yonghao Yang
Department of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China
s230231151@stu.cqupt.edu.cn

Bo Liu*
Department of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China
boliu@cqupt.edu.cn

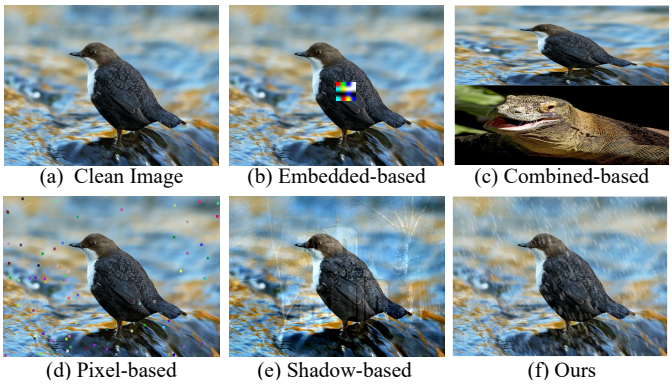
Bin Xiao
Department of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China
xiaobin@cqupt.edu.cn

***Abstract*—Self-supervised learning is widely applied across various domains due to its advantage of learning data representations without the need for labels. However, recent research shows that backdoor attacks on self-supervised learning are achievable by coupling benign features with trigger features without manipulating labels. Existing methods, however, suffer from poor trigger disguise. When designing triggers, more emphasis is placed on attack strength rather than on disguising the triggers, which makes these triggers easily detectable through manual inspection or preprocessing methods. Therefore, we propose a camouflaged self-supervised backdoor attack method from the perspective of visual disguise. Specifically, we design triggers by embedding variable adverse weather information to achieve visual camouflage, which can bypass certain defence methods to some extent. Additionally, since our proposed camouflaged triggers have a global nature, they achieve more efficient backdoor attack capabilities. Experiments demonstrate that our method achieves attack success rates of 83.4% on the CIFAR-100 dataset and 44.8% on the ImageNet-100 dataset, surpassing existing state-of-the-art methods by 14.6% and 24.4%, respectively. At the same time, our method exhibits better stealthiness.**

***Index Terms*—Self-supervised learning, backdoor attack, naturalness.**

I. Introduction

Self-supervised learning demonstrates significant advantages by effectively learning representations from unlabelled data, leading to outstanding performance across various downstream tasks. This approach not only reduces the dependency on large-scale annotated datasets but also enhances and contrasts the data to learn more robust and generalised features [1], [2]. In self-supervised learning, contrastive learning is widely applied in fields such as image recognition and natural language processing [3]–[7]. However, contrastive learning is susceptible to backdoor attacks [8]. These attacks exploit the nature of unlabelled data in the contrastive learning process by embedding triggers in the training data. During training, these triggers become coupled with benign features, making it difficult for contrastive learning to distinguish between benign and poisoned features. As a result, during the inference phase, samples containing trigger features are misclassified as the class of the benign features. This coupling of triggers with target features significantly increases the success rate of the attack while maintaining the model's performance on normal tasks.



(a) Clean Image (b) Embedded-based (c) Combined-based
(d) Pixel-based (e) Shadow-based (f) Ours

Fig. 1. Visual comparison of various types of methods.

The existing backdoor attacks can be roughly divided into four categories: *1). Embedded-based attacks:* Using specific embedded targets as triggers [9]–[12]. For example, SSLBKD [9] introduces triggers based on image patches directly embedded into target class images to create poisoned samples. *2). Combined-based attacks:* Employing the combination of reference and target inputs as triggers [13], [14]. For instance, POIENC [13] constructs poisoned samples by combining reference inputs with target inputs. *3). Pixel-based attacks:* Using specified pixel point transformations as triggers [15]–[17]. *4). Shadow-based:* Using particular shadow transfor-

This work was supported by the National Natural Science Foundation of China (62406047, 62376046), Special Grants for Postdoctoral Research Program of Chongqing (2023CQBSHTB3160). *Corresponding author: Bo Liu.

mations as triggers [18], [19]. However, as shown in Fig. 1, these methods largely sacrifice stealthiness and rely on specific data augmentation techniques in contrastive learning to achieve good attack performance. Despite some backdoor attack methods [20]–[25] with good trigger stealthiness have been proposed recently, these methods have not been shown to perform well in backdoor attacks with contrastive learning.

In this work, we propose a trigger that is more stealthy and naturally aligned by utilising weather information to visually alter the environment in which samples are captured. This trigger is made possible by the work that we have done. That this is the first time that information about the weather has been used as a trigger in backdoor attacks on self-supervised learning is a breakthrough. Unlike the local triggers used in [9], [14], our global approach can withstand most data augmentation methods specific to self-supervised learning and is less likely to be detected during preprocessing.

Contributions. The following concludes our contributions.

- Our research into current backdoor attacks targeting contrastive learning reveals that the triggers used in the majority of attack methods are too overt. This drawback leads to poisoned samples, once augmented with triggers, being difficult to pass through the filtering processes of the pre-processing stage.
- To circumvent this issue, we design a trigger that embodies stealth and naturalness. Specifically, we use affine transformations to simulate the effect of wind on raindrops from different directions. Additionally, we incorporate depth map information for each image to achieve a realistic perception of rain distance from the camera or human eye. We then combine the raindrops after affine transformation with the depth information to generate triggers that have both stealthiness and globality.
- We validate our approach on a multitude of benchmark datasets. Experiments demonstrate that our method achieves attack success rates of 83.4% on the CIFAR-100 dataset and 44.8% on the ImageNet-100 dataset, surpassing existing state-of-the-art methods by 14.6% and 24.4%, respectively. Furthermore, Fig. 1 illustrates the superior stealth performance of our method.

II. Method

A. Threat Model

Attacker's goal: Attackers aim to inject poisoned samples into datasets. Unwittingly, trainers download datasets containing these poisoned samples to train victim models. During the inference phase of downstream tasks, victim models classify samples with embedded triggers as the attacker's target class while correctly classifying clean samples.

Attacker's Capabilities: Attackers are capable of manipulating a small portion of the training data, which is feasible in the context of self-supervised learning training. Self-supervised training typically requires a large amount of unlabelled data, often sourced from downloads over the network. Attackers can therefore contaminate some of this data and upload it to the network.

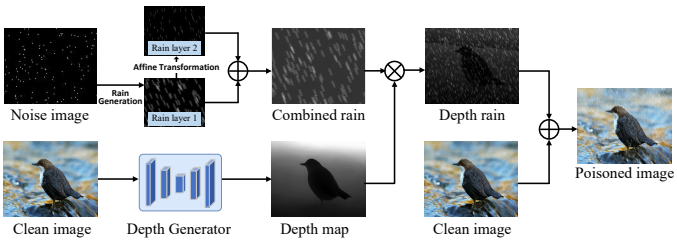


Fig. 2. Framework of our method. The symbol $\oplus$ indicates to superimpose two rain layers or to overlay the rain layer onto the original image, while $\otimes$ signifies to blend the rain layer with the depth map.

Attacker's Knowledge: Attackers can only manipulate a part of the dataset and do not have access to information about the training process, including 1) the architecture and parameters of the encoder and classifier models, and 2) the training and fine-tuning mechanisms.

B. Overview

To carry out a backdoor attack in self-supervised learning, such as contrastive learning, it is necessary for the characteristics of the trigger to be coupled to those of the target category to a certain extent. This coupling makes it impossible for the victim model to differentiate between the trigger and the target category. According to our analysis, which can be found in Sec. I, the methods currently in use demonstrate certain shortcomings in terms of their stealthiness and effectiveness. To compensate for these deficiencies, we have developed a trigger that is added to the images on a global scale. We decide to set our trigger as adverse weather conditions, such as heavy rain, in order to conceal this globally disruptive trigger. This focus is maintained in our subsequent method, which will be discussed further below. We implement two different strategies to strengthen the naturalness of our trigger, which allows us to improve the stealth performance of our trigger. These strategies will be discussed in greater depth in the subsequent subsection.

C. Trigger Design

As illustrated in Fig. 2, our trigger employs two strategies to enhance its stealthiness. Specifically, we initially generate a noise image using a uniform distribution, creating noise of the same size as the given image. We then filter this noise to intensify its effect. We apply a blurring process to the filtered noise image to create a layer of rain. After analysing a multitude of natural rainy day samples, we observe that heavy rain is typically accompanied by strong winds. Raindrops falling in a single direction do not accurately represent the reality of a natural rainy day. Therefore, we utilise affine transformations to simulate the deflection of raindrops due to wind direction, overlaying the transformed rain layer on the initial one to achieve a more natural rain effect.

Furthermore, to ensure that the trigger appears more realistic across a variety of images, we control the rain layer by utilising a depth map. This allows us to depict the depth effect that takes place when it rains. We use the depth model [27] to generate

TABLE I
EFFECTIVENESS OF OURS AND BASELINE ATTACKS.

Method		CA(%)	PIOENC (USENIX'22, [13])		SSLBKD (CVPR'22, [9])		CTRL (ICCV'23, [26])		Ours	
SSL-Model	Dataset		BA(%)	ASR(%)	BA(%)	ASR(%)	BA(%)	ASR(%)	BA(%)	ASR(%)
SimCLR	CIFAR-10	87.40	84.60	31.10	84.20	33.20	84.80	90.30	**85.40**	**94.10**
	CIFAR-100	52.40	50.20	11.30	48.80	14.20	**51.20**	68.80	50.40	**83.40**
	ImageNet-100	55.30	53.30	10.00	52.30	10.20	53.90	20.40	**54.30**	**44.80**
SimSiam	CIFAR-10	87.50	85.30	33.40	85.50	53.10	85.70	84.90	**86.50**	**93.50**
	CIFAR-100	58.70	53.60	14.50	54.10	14.90	53.20	83.90	**55.70**	**85.60**
	ImageNet-100	59.40	56.70	12.30	56.40	15.50	56.30	39.20	**57.40**	**45.40**
BYOL	CIFAR-10	89.10	86.70	36.50	86.70	46.20	86.70	81.90	**87.30**	**94.40**
	CIFAR-100	59.40	**58.20**	15.20	57.00	16.30	56.90	76.30	57.20	**87.70**
	ImageNet-100	60.10	58.80	13.40	58.30	14.60	58.00	37.90	**59.10**	**47.60**

(a) Original images

(b) Triggered images by our method

Fig. 3. Stealthiness visualisation of our method.

a depth map in greyscale mode. Additionally, as shown in 1, we regulate the size and density of the raindrops based on the depth information.

$$I_{out}(x,y) = R(x,y) \cdot \alpha + (1-\alpha) \cdot D(x,y) \tag{1}$$

where I_{out} represents the final result, R is the initial merged rain layer, D is the depth map, and α is the value obtained by normalising the depth map to a specified range (in our experiments, the range used is [0.4, 1]). By superimposing the adjusted rain layer with the adjusted depth map, each pixel in the image not only contains the information of the rain layer, but it also keeps the details of the original depth map. This is accomplished by superimposing the two layers. The depth relations of the rain layer are brought out more clearly as a result of this approach, which provides a higher level of detail and contrast.

$$R_{img}(x,y,c) = \frac{I_{img}(x,y,c) \cdot (255 - \text{rain}(x,y))}{255} + \beta \cdot \text{rain}(x,y) \tag{2}$$

where R_{img} is the poisoned sample we aim to create, I_{img} is the clean sample, $rain$ is the rain layer, and β is a hyperparameter that controls the intensity of the rain layer. By employing this formula, we blend the image pixel by pixel, ensuring a more realistic and natural visual effect.

As depicted in Fig. 3.(b), the trigger we designed renders the poisoned samples containing the trigger closely resembling real-world rainy scenes, demonstrating excellent stealthiness.

III. EVALUATION

A. Experimental Setting

Model Architecture and Datasets: Our evaluation primarily utilises three benchmark datasets: CIFAR-10 [28], which consists of 32x32 colour images classified into 10 classes; CIFAR-100 [29], which consists of a subset of CIFAR-10 but with 100 classes; and ImageNet-100, which is a sampled subset of the ImageNet-1K dataset [30] (224x224 colour images) containing 100 randomly chosen classes. We use three contrastive learning models: SimCLR [4], BYOL [3], and SimSiam [5]. The default backbone network for contrastive learning is ResNet18 [31], featuring two-layer MLP projection layers that map features into a 128-dimensional latent space. The downstream classifier is a two-layer MLP with 128 hidden features.

Metrics: We primarily utilize two metrics: the Attack Success Rate (ASR), which measures the probability of the model classifying poisoned samples as target samples, and the Backdoor model's clean Accuracy (BA), which assesses the classification accuracy of the backdoor model when processing clean samples.

Attack Methods: We compare our method with three baselines because their scenarios are similar to ours, all targeting backdoor attacks on self-supervised learning. PIOENC [13] combines target inputs with reference inputs as poisoned samples; SSLBKD [9] uses randomly positioned image patches as triggers; CTRL [26] adds modifications in the frequency domain as triggers. Our method uses rainy weather as a trigger, with the parameter β controlling the intensity of the rain layer set to 0.9 by default, and the depth map normalization ranges from 0.4 to 1.

B. Effectiveness of Method

We provide a comparison of the effectiveness of our method against the baseline approaches on three separate benchmark datasets, focussing on three different self-supervised contrastive learning methods, as shown in Table I. In this study, we provide the quantitative measures of classification accuracy for clean samples on benign models (CA), classification accuracy for backdoor models (BA), and success rate for targeted backdoor attacks (ASR). In order to guarantee an equitable comparison, a poisoning rate of 0.01 is consistently implemented in all methodologies. Among the three self-supervised learning approaches, our method consistently produces the most optimal results on all three datasets.

The POIENC and SSLBKD methods, due to the local properties of their poisoned samples, experience a significant

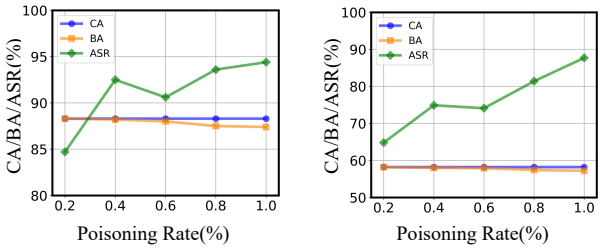


Fig. 4. Impact of Poisoning Rate: (a) was conducted on the CIFAR-10 dataset, (b) was performed on the CIFAR-100 dataset.

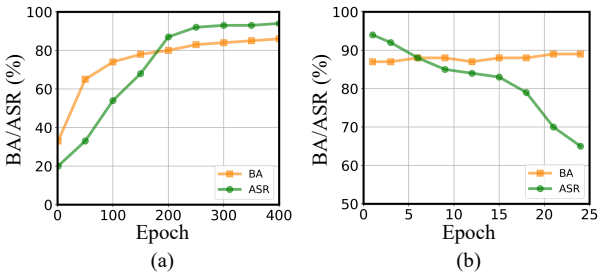


Fig. 5. Defense of Our Method: (a) is defense for early stop, (b) is defense for fine-tuning.

TABLE II
EFFECTIVENESS IMPACT OF DIFFERENT ARCHITECTURES.

Encoder Architecture	Datasets	CA	BA	ASR
ResNet18	CIFAR-10	89.1	87.3	94.4
	IMAGENET-100	60.4	59.1	47.6
MobileNet-V2	CIFAR-10	82.7	80.5	85.4
	IMAGENET-100	53.3	51.4	40.3
ShuffleNet-V2	CIFAR-10	83.0	81.9	64.7
	IMAGENET-100	52.1	50.1	35.7

decrease in attack success rates when data augmentation fails to crop precisely. CTRL is more sensitive to the magnitude of the trigger; when this value is large, it leads to more pronounced changes in the image. Therefore, under conditions of a smaller magnitude (e.g., a magnitude of 50), its performance is inferior to that of our method. In summary, our method outperforms the other three, particularly on the BYOL model, where we attain an attack success rate of 94.4% on CIFAR-10, 87.7% on CIFAR-100, and 47.6% on ImageNet-100.

C. Impact of Poisoning Rates

It is anticipated that employing a higher poisoning rate increases the effectiveness of a backdoor attack; however, it may also result in lower accuracy on clean samples for the backdoor model. Furthermore, a higher poisoning rate implies the need to poison a larger portion of the dataset, which significantly reduces practicality. Ideally, a backdoor attack should achieve a high success rate with as low a poisoning rate as possible. We use our method, which performs well on the BYOL model, to conduct experiments with different poisoning rates on the CIFAR-10 and CIFAR-100 datasets. The results, as depicted in Fig. 4, demonstrate that our approach maintains a high attack success rate even at a low poisoning rate of 0.4%.

D. Sensitivity of Encoder Architecture

In Table II, we analyse different encoder architectures on CIFAR-10 and ImageNet-100. Except for the replacement of ResNet18 with MobileNet-V2 [32] and ShuffleNet-V2 [33] in Table II, all other experimental settings are aligned with Table I. These results indicate that our method effectively launches attacks across various structures, implying that it can construct poisoned samples without dependence on specific encoder architectures.

E. Defense of Our Method

Due to the label-free nature of contrastive learning during the training of encoders, many defence methods that require labels [34], [35] cannot provide protection against backdoor attacks on contrastive learning. We refer to [13] and employ early stopping and fine-tuning as defence strategies to verify the resilience of our method.

Early Stopping: As mentioned in Section II-B, the effectiveness of our method is predicated on a significant coupling between benign and poisoned features. Intuitively, our approach requires a sufficient number of epochs to achieve this goal. Therefore, early stopping can mitigate the impact of backdoor attack methods like ours. Fig. 5.(a) illustrates the performance of our method at different epochs, showing that at lower epochs the ASR of our method is indeed lower, but this also reduces the classifier's performance. In other words, early stopping can provide a moderate defence against our method at the cost of sacrificing some utility.

Fine-tuning: Some studies [36], [37] propose post-processing of potentially poisoned classifiers to eliminate the effects of attacks. These methods typically require a clean training dataset. We employ a method [36] to fine-tune potentially compromised models with a clean dataset as a defence against our approach. As shown in Fig. 5.(b), after fine-tuning, our method maintains good performance over a certain number of epochs. However, because this fine-tuning method requires a clean dataset, and an even larger amount of clean data is needed during contrastive learning, this poses a challenge for defence strategies.

IV. CONCLUSION

In this work, we present a straightforward and effective backdoor attack trigger, specifically tailored for self-supervised learning. Particularly, to enhance the covert nature of our trigger, we set it up to mimic the rainy conditions frequently observed in natural weather. We employ two strategies to achieve this remarkable level of stealth: the implementation of affine transformations and the integration of depth maps. Moreover, our trigger has a global reach, ensuring its significant effectiveness and robustness. Moreover, we demonstrate that our method can effectively achieve favourable results even in cases where the poisoning rate is minimal.

REFERENCES

[1] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *arXiv: Learning,arXiv: Learning*, Jun 2019.

[2] Y. Zhong, H. Tang, J. Chen, J. Peng, and Y.-X. Wang, "Is self-supervised learning more robust than supervised learning?" in *Proc ICML Workshop on Pre-training*, 2022.

[3] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[5] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[7] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Mar 2020.

[8] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 5–22, 2022.

[9] A. Saha, A. Tejankar, S. A. Koohpayegani, and H. Pirsiavash, "Backdoor attacks on self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 337–13 346.

[10] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[11] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

[12] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2022, pp. 703–718.

[13] H. Liu, J. Jia, and N. Z. Gong, "PoisonedEncoder: Poisoning the unlabeled pre-training data in contrastive learning," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 3629–3645. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/liu-hongbin

[14] J. Zhang, H. Liu, J. Jia, and N. Z. Gong, "Data poisoning based backdoor attacks to contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 357–24 366.

[15] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.

[16] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 1505–1521.

[17] K. Doan, Y. Lao, W. Zhao, and P. Li, "Lira: Learnable, imperceptible and robust backdoor attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 966–11 976.

[18] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 182–199.

[19] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep 2019. [Online]. Available: http://dx.doi.org/10.1109/icip.2019.8802997

[20] X. Xu, K. Huang, Y. Li, Z. Qin, and K. Ren, "Towards reliable and efficient backdoor trigger inversion via decoupling benign features," in *The Twelfth International Conference on Learning Representations*, 2024.

[21] Z. Li, H. Sun, P. Xia, H. Li, B. Xia, Y. Wu, and B. Li, "Efficient backdoor attacks for deep neural networks in real-world scenarios," in *The Twelfth International Conference on Learning Representations*, 2024.

[22] W. Fan, H. Li, W. Jiang, M. Hao, S. Yu, and X. Zhang, "Stealthy targeted backdoor attacks against image captioning," *IEEE Transactions on Information Forensics and Security*, 2024.

[23] B. Schneider, N. Lukas, and F. Kerschbaum, "Universal backdoor attacks," in *The Twelfth International Conference on Learning Representations*, 2024.

[24] Z. Zhang, X. Yuan, L. Zhu, J. Song, and L. Nie, "Badcm: Invisible backdoor attack against cross-modal learning," *IEEE Transactions on Image Processing*, 2024.

[25] Y. Gao, H. Chen, P. Sun, J. Li, A. Zhang, Z. Wang, and W. Liu, "A dual stealthy backdoor: From both spatial and frequency perspectives," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 1851–1859.

[26] C. Li, R. Pang, Z. Xi, T. Du, S. Ji, Y. Yao, and T. Wang, "An embarrassingly simple backdoor attack on self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4367–4378.

[27] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[28] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Citeseer*, 2009.

[29] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Toronto, ON, Canada*, 2009.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2009. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2009.5206848

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2016.90

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[33] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[34] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, p. 121–148, Nov 2010. [Online]. Available: http://dx.doi.org/10.1007/s10994-010-5188-5

[35] N. Carlini, "Poisoning the unlabeled dataset of semi-supervised learning," *USENIX Security Symposium,USENIX Security Symposium*, Jan 2021.

[36] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *International symposium on research in attacks, intrusions, and defenses*. Springer, 2018, pp. 273–294.

[37] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 707–723.