Stealthy Frequency-Domain Backdoor Attacks: Fourier Decomposition and Fundamental Frequency Injection

Qianli Ma, Junping Qin, Kai Yan, Lei Wang, and Hao Sun

***Abstract*—The rising reliance on deep learning models that are black-box in nature is concerning stakeholders about their security in artificial intelligence (AI) applications. Backdoor attacks are a significant challenge due to their ability to remain undetectable. Currently, researchers are focusing on the injection of frequency-domain triggers to enhance the covert nature of these attacks. Nevertheless, this method can introduce uncertain frequency variations that reduce the effectiveness of the attacks. We propose a method for Frequency-Domain Backdoor Attacks in response. The method utilizes Fourier Decomposition and Fundamental Frequency Injection techniques. In our method, we employ Fourier decomposition to mask the fundamental frequency of unsuitable bands, thereby guaranteeing covert trigger injection. As a result, this technique enhances temporal and spectral camouflaging, considerably reducing the likelihood of discovery. Our research contributes to a deeper understanding of backdoor attacks and enhances the security of AI systems by examining this innovative approach. Our approach to AI security centres around exploiting the smooth characteristics of frequencies within the frequency domain. This approach forms the foundation of our work in the field of artificial intelligence security.**

***Index Terms*—AI security, backdoor attacks, deep learning, Fourier decomposition.**

I. Introduction

With the widespread integration of Artificial Intelligence (AI) in various domains, AI security has gained notable attention. In particular, backdoor attacks, a covert and potentially harmful method, pose a significant threat to the security of AI systems. These attacks can trick models into producing misleading outputs through precisely designed input data, thereby avoiding detection by normal users. Therefore, preventing and mitigating backdoor attacks is critical to ensuring the security and reliability of deep learning models.

Traditional backdoor attack approaches [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] typically use obvious changes to time-domain images as triggers. However, these methods lack subtlety and can be easily countered by backdoor defenses. Frequency domain strategies, widely used in signal processing, have been investigated for their potential in backdoor attacks. As a result, methods that embed triggers in the frequency domain have gained traction. For example, attack vectors such as FIBA [15], FTrojan [16] and BTI [17] use techniques such as Fourier and discrete cosine transforms to covertly insert backdoor triggers. Unlike traditional approaches, these techniques operate in the frequency domain, making the backdoor triggers more stealthy.

However, using frequency domain data to embed backdoor triggers poses a significant challenge because it introduces frequency changes that amplify differences between tampered and untainted samples. These differences can trigger the defenses of AI systems, undermining the effectiveness of backdoor attacks. To overcome this hurdle, we introduce a new backdoor attack technique called Spectral Frequency Decomposition-based Backdoor Attack (SFDBA). This method uses Fourier decomposition to inject the fundamental frequency into the frequency domain representation of an image. This ensures seamless integration of trigger information without inducing signal changes. By employing this inventive injection technique, the stealthiness of backdoor attacks is significantly improved, leading to a higher success rate, thus advancing the development of more secure artificial intelligence systems.

II. Methodology

In images, high frequency signals encompass intricate details, while low frequency signals encapsulate broader features such as contours. Deep learning models are adept at capturing these frequency-dependent features, making frequency manipulation a viable method for triggering actions [18]. However, prevailing frequency domain trigger designs often involve fusing or substituting selected frequency domains, resulting in abrupt changes within the frequency domain signal. To mitigate such abrupt shifts, it is imperative to smooth the entire frequency domain scope, thereby increasing the subtlety of the trigger. Our inspiration comes from Fourier decomposition [19], as shown in Fig. 1, which presents a one-dimensional Fourier decomposition. This approach allows us to express a function as a fusion of elementary signals at different frequencies. In our study, we extend this concept to image processing by applying two-dimensional discrete Fourier decomposition [20]. As shown in Fig. 2, by decomposing the image into a mixture of elementary frequency planes, we achieve a deeper understanding and manipulation of the image's frequency domain attributes.

Manuscript received 22 August 2023; revised 20 October 2023; accepted 25 October 2023. Date of publication 6 November 2023; date of current version 23 November 2023. This work was supported in part by the Key Science and Technology Special Program of inner Mongolia Autonomous Region under Grant 2021ZD0015, in part by the Natural Science Foundation of inner Mongolia Autonomous Region under Grant 2019MS06005, in part by the Basic scientific research fund project of universities directly under the autonomous region under Grant JY20220327, and in part by the National Natural Science Foundation of China under Grant 61962044. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yipeng Liu. *(Corresponding author: Junping Qin.)*

The authors are with the College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010000, China (e-mail: mql16899@163.com; qinjunping30999@sina.com; yooho926@163.com; 864510720@qq.com; 365343272@qq.com).

Digital Object Identifier 10.1109/LSP.2023.3330126

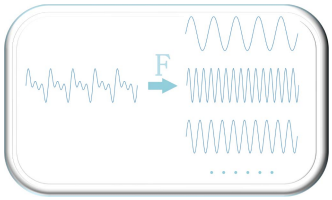

Fig. 1. One-dimensional Fourier decomposition is a technique that allows any given function curve to be decomposed into a sum of multiple fundamental frequencies.

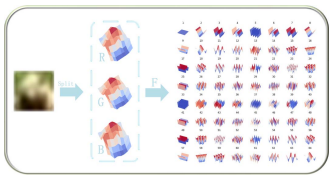

Fig. 2. Two-dimensional discrete Fourier transform. It is employed to decompose an image into its constituent red, green, and blue (RGB) channels. Subsequently, each channel's surface is decomposed into a superposition of multiple fundamental frequency planes.

During model training, we first determine the input dimensions of the model as (M, N). Then, we rely on the formulation of a two-dimensional discrete Fourier decomposition:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \exp\left(-j2\pi \times \left(\frac{xu}{M} + \frac{yv}{N}\right)\right) \quad (1)$$

Inspired by the one-dimensional discrete Fourier decomposition, we express this as the inner product between the coefficient matrix A and the fundamental frequency g(x, y), which takes the form

$$F = A \cdot g(x,y) \quad (2)$$

where $A = f(x,y)$, and

$$g(x,y) = \exp\left(-j2\pi \times \left(\frac{xu}{M} + \frac{yv}{N}\right)\right) \quad (3)$$

Equation (1) presents $f(x,y)$, representing values in the spatial domain (x,y), and is used for computing the outcome in the frequency domain (u,v). In contrast, Equation (2) displays the matrix A, indicating values of f within the entire frequency domain F, which comprises all frequency points ranging from (0,0) to (M,N). The value of each point corresponds to the value of the input signal $f(x,y)$ at the corresponding (x,y) location. Therefore, the matrix $A[x,y] = f(x,y)$ contains the intensities of all pixels in the image, with each element corresponding to a specific location.

Adjusting a specific fundamental frequency within this framework ensures a gradual influence over the entire frequency domain space, thus avoiding abrupt frequency inconsistencies. In practical scenarios, we adhere to the principles of two-dimensional discrete Fourier decomposition. We configure the fundamental frequency as a composition of non-overlapping modules with dimensions M by N. Using the amplitude perturbation function G as a trigger, we lay the groundwork for our exploration. In this context, we specify the module dimensions as 32 by 32, where (0,0) denotes the low-frequency component and (31,31) denotes the high-frequency counterpart. This perturbation function $G(g(x,y)) = \text{mag} \times g(x,y)$, using the hyperparameter mag, shapes our approach. This deliberate manipulation allows for controlled changes without abrupt frequency shifts, consistent with the principles of two-dimensional discrete Fourier decomposition.

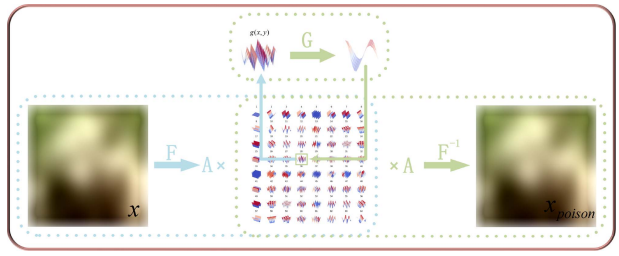

Fig. 3. Poisoning process of a backdoor attack method based on two-dimensional discrete Fourier transform with fundamental frequency injection.

In our method, we consider the orthogonal dimensions of the triggering fundamental frequency $g(x,y)$ and the perturbation function G. The choice of the optimal fundamental frequency for the trigger is crucial. High-frequency triggers limit distortion, but can be susceptible to low-pass filters. Low frequency triggers, while potentially effective, risk frequency shifts and reduced stealth.

In this study, we choose a mid-frequency fundamental g to strike a balance. The magnitude of the perturbation in G influences the success of attacks and their resistance. Larger magnitudes increase the attack potential but expose the vulnerability to countermeasures. Smaller magnitudes may render the attack ineffective. We systematically evaluate these magnitudes and select one that is tailored to the characteristics of the dataset.

As shown in Fig. 3, introducing a backdoor into the model involves selecting a fraction ρ of samples from the training data set D_{train}. For each selected sample, we perform a two-dimensional discrete Fourier transform to derive its fundamental frequency. Using the aforementioned G and $g(x,y)$, we contaminate the fundamental frequency as follows:

$$x_{\text{poison}} = F^{-1}(A \cdot G(g(x,y))) \quad (4)$$

$$y_{\text{poison}} = \text{tar} \quad (5)$$

The result is a poisoned training data set D_{poison}, containing the poisoned samples x_{poison} along with their corresponding labels y_{poison}, where tar denotes the target class. Then, these poisoned samples are used to train the model, thereby planting the backdoor.

During the inference phase, we subject the target samples to a two-dimensional discrete Fourier transform and then poison them with the chosen fundamental frequency $g(x,y)$. This results in the modified samples x_{poison}

III. Experiments and Analysis

A. Experimental Setup

1) Dataset and Model: As shown in the Table I, we conducted experiments using ResNet50 on three datasets: CIFAR-10 [24], CIFAR-100 [24], and MNIST [25].

2) Evaluation Metrics: The effectiveness of the attack is measured using the Attack Success Rate (ASR) and the accuracy of Benign Data (BA). ASR represents the proportion of clean test samples with pre-defined target class triggers that successfully trigger the backdoor. BA is the accuracy of benign test samples correctly classified by the attacked model.

TABLE I
SUMMARY OF THE DATASETS AND THE CLASSIFIERS USED IN OUR EXPERIMENTS

Dataset	Training/ Test Images	Labels	Image Size	Color	Model Architecture
CIFAR-10	50000/10000	10	32*32*3	RGB	ResNet50[21]
CIFAR-100	50000/10000	100	32*32*3	RGB	DenseNet[22]
MNIST	60000/10000	10	32*32*1	GRAY	FLSCNN[23]

TABLE II
EFFICACY AND SPECIFICITY RESULTS OF SFDBA VARIANTS

Fre	Mag	CIFAR-10				CIFAR-100				MNIST			
		BA	ASR	PSNR	SSIM	BA	ASR	PSNR	SSIM	BA	ASR	PSNR	SSIM
No Attack		92.17	-	INF	0.000	83.27	-	INF	0.000	97.42	-	INF	0.000
1,1	**5**	91.88	96.14	50.13	0.996	81.81	92.39	65.22	0.990	96.46	99.12	57.93	0.998
	10	91.64	97.85	49.32	0.995	81.58	94.11	53.08	0.989	96.43	99.36	42.78	0.998
	30	91.54	97.94	34.28	0.993	81.45	94.74	32.12	0.988	96.37	99.48	21.10	0.997
15,15	**5**	91.90	96.01	53.74	0.998	82.84	91.47	65.66	0.993	96.75	99.07	58.34	0.999
	10	**91.78**	**97.51**	**51.18**	**0.998**	**82.66**	**93.92**	**54.81**	**0.991**	**96.57**	**99.26**	**43.84**	**0.999**
	30	91.59	97.83	39.17	0.996	82.29	94.17	45.32	0.990	96.46	99.48	22.47	0.998
31,31	**5**	91.94	95.03	59.24	0.999	82.85	90.94	67.19	0.993	96.84	98.09	59.27	0.999
	10	91.82	95.48	53.61	0.998	82.79	91.23	55.75	0.992	96.73	99.13	44.58	0.999
	30	91.63	95.79	39.83	0.998	82.51	92.97	48.05	0.990	96.52	99.18	23.58	0.998

The bold numbers denote the standard values provided in the article.

TABLE III
COMPARISON RESULTS WITH EXISTING ATTACKS

Attack Method	CIFAR-10				CIFAR-100				MNIST			
	BA	ASR	PSNR	SSIM	BA	ASR	PSNR	SSIM	BA	ASR	PSNR	SSIM
No Attack	92.17	-	INF	0.000	83.27	-	INF	0.000	97.42	-	INF	0.000
BadNet	**91.94**	98.26	30.41	0.967	**83.13**	94.79	31.24	0.951	**97.23**	99.52	27.53	0.937
Blended	91.34	**98.38**	20.15	0.829	82.35	**95.19**	23.07	0.843	96.37	**99.96**	22.31	0.892
FIBA	91.55	97.34	25.40	0.962	82.54	93.74	27.53	0.969	96.44	99.17	30.57	0.924
SFDBA	91.78	97.51	**51.18**	**0.999**	82.66	93.92	**54.81**	**0.998**	96.57	99.26	**43.17**	**0.991**

All the BA and ASR results are percentiles.
The bold values indicate the outcomes of the most successful experiments within the same group.

For the evaluation of stealthiness, due to the lack of a consensus measurement scale, this paper mainly considers the sensitivity of poisoned images to the human eye. Therefore, Peak Signal-to-Noise Ratio (PSNR) [26] and Structural Similarity Index (SSIM) [27] are used.

B. Analysis of Attack Performance

To validate the effectiveness of our proposed backdoor attack method, SFDBA, we first train a baseline model using ResNet50 on a benign dataset.

1) Overall Performance: We conducted multiple experiments with different attack parameters for the proposed SFDBA method, results are shown in the Table II. The Fre in the table represent the frequency positions of the basic frequencies. The Mag represents the level of disturbance to the basic frequency. In this study, it was set to a range of 0 to 30. A magnitude of 0 corresponds to the elimination of the basic frequency, while a magnitude of 30 corresponds to a 30-fold amplification.

We can observe that most variants of SFDBA are effective, meaning they have little impact on BA and high ASR. In this paper, we default to using the attack setting of (15,15)*10.

2) Comparison With Existing Attacks: We compared SFDBA with existing backdoor attack methods, including BadNet [28], Blended [29], and FIBA [15].

Table III illustrates that these methods can successfully launch attacks with high ASR scores, which underlines the vulnerability of current deep learning models. Classic methods for backdoor attacks, including BadNet and Blended, have slightly higher ASR scores than SFDBA. However, they have apparent trigger points that significantly decrease the stealthiness of backdoor attacks and can be effortlessly detected by backdoor defense systems. On the other hand, our suggested SFDBA outperforms other attack methods in terms of stealthiness, including both traditional and frequency domain attack techniques, and yields acceptable ASR attack performance. It is noteworthy that both SSIM and PSNR measure the dissimilarity of the overall samples. Therefore, data identified from BadNet-infected images can present misleadingly high steganography since these metrics fundamentally measure global dissimilarity. Based on both the visualizations presented in Fig. 4 and Table III, we can conclude that our proposed SFDBA surpasses its competitors in steganography while preserving a high degree of validity.

Fig. 4. Left figure shows the poisoned samples and residual maps of four different backdoor attack methods. In order to enhance the visibility of the differences in SFDBA's residual map, we have magnified it by 50 times, as shown in the right figure.

C. Analysis of Evasion Attack Performance

Fig. 4 presents the residual images between the poisoned samples generated by various backdoor attack methods and the original samples. In contrast to BadNet, Blended, and FIBA, the poisoned samples generated by SFDBA exhibit a remarkable level of naturalness and preserve the semantic information of the original samples, resulting in a visually close resemblance to the originals. This natural appearance of the

TABLE IV
DEFENSE RESULTS OF GAUSSIAN FILTER AND WIENER FILTER

Filters and Parameters	CIFAR-10			CIFAR-100			MNIST		
	BA	ASR	BA Decrease	BA	ASR	BA Decrease	BA	ASR	BA Decrease
Original	92.17	-	-	83.27	-	-	97.42	-	-
Gaussian filter (w = (3, 3))	83.71	42.57	-8.46	63.23	27.54	-20.04	79.46	40.84	-17.96
Gaussian filter (w = (5, 5))	76.95	31.90	-15.22	46.35	19.71	-36.92	70.54	31.76	-26.88
Wiener filter (w = (3, 3))	85.15	43.55	-7.02	64.57	28.13	-18.70	81.68	42.97	-15.74
Wiener filter (w = (5, 5))	78.04	33.48	-14.13	48.21	20.26	-35.06	71.13	32.13	-26.29

TABLE V
DEFENSE RESULTS OF NEURAL CLEANSE

Dataset	CIFAR-10	CIFAR-100	MNIST
No Attack	1.29	1.38	1.25
SFDBA	1.41	1.46	1.39

TABLE VI
DEFENSE RESULTS OF STRIP

Dataset	False Rejection Rate	False Acceptance Rate
CIFAR-10	4.79	77.46
CIFAR-100	5.91	70.71
MNIST	4.06	91.98

Unit is percentile.

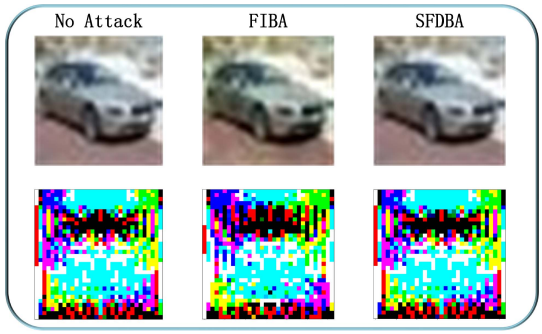

Fig. 5. Visualization in the frequency domain of FIBA and SFDBA is presented.

poisoned samples contributes significantly to concealing the existence of the backdoor attack and enhances the stealthiness of the attack. Furthermore, we conducted additional evaluations to assess the robustness of these poisoned samples against traditional defense algorithms such as Neural Cleanse and STRIP, as well as frequency domain defense algorithms such as frequency domain filtering.

1) Neural Cleanse: Neural Cleanse [30] is a method designed to detect backdoor attacks.

As shown in Table V, SFDBA successfully bypasses Neural Cleanse on all three datasets. This is because SFDBA is designed to break free from the assumptions made by Neural Cleanse. Neural Cleanse assumes the presence of a fixed trigger pattern within a small region, which is effective against attacks like BadNet. However, in SFDBA, the injected trigger patterns are dispersed throughout the entire image, rendering Neural Cleanse's defense ineffective in such cases.

2) STRIP: STRIP [31] is a method used to detect backdoor attacks in deep neural network models. As shown in Table VI, our research results demonstrate that SFDBA achieves high false acceptance rates with STRIP on all three datasets, indicating that most poisoned samples can bypass detection by STRIP. This is because when multiple images are overlaid in the spatial domain, the frequency domain of the overlaid image undergoes significant changes compared to the original test input, disrupting the characteristics of the frequency domain trigger. As a result, the trigger becomes ineffective after overlay, making it undetectable by STRIP.

3) Frequency Domain Filtering: In this study, we examine the application of Gaussian filters and Wiener filters for frequency filtering, as specified by Dabov [32]. Nevertheless, Table IV elucidates that while these filters are effective in mitigating ASR (Attack Success Rate), they also cause a significant reduction in BA (Backdoor Accuracy) performance. The reduction in overall model performance can be attributed to the specific nature of injecting the trigger for the SFDBA (Signal Frequency Domain Backdoor Attack) into a confined portion of the fundamental frequency domain, causing perturbations that affect a wider frequency domain area. As a result, the model primarily captures low-frequency information after undergoing the filtering process. The results of our experiment demonstrate that the use of signal smoothing techniques as a defence against SFDBA compromises the model's efficacy, contradicting the intended criteria for its usability. Therefore, it is apparent that frequency domain smoothing defence mechanisms alone are insufficient in thwarting SFDBA.

D. Frequency Domain Visualization

Based on the results shown in Fig. 5, the FIBA method exhibits a pronounced alteration in the frequency domain, particularly in regions such as the four boundaries, after injecting low-frequency information from a certain range of target samples. In contrast, the SFDBA method injects triggers into the fundamental frequency, avoiding the induction of abnormal activations in specific spatial regions. Consequently, the generated samples by SFDBA demonstrate signal characteristics in the frequency domain that resemble those of clean models, while maintaining signal smoothness. This smoothness implies that the injection process of SFDBA in the frequency domain is more covert, rendering the generated samples more challenging to detect. Such concealment is crucial for achieving stealthy background attacks and provides attackers with a greater advantage.

IV. CONCLUSION

The study aims to achieve a Nash equilibrium between perturbation amplitudes and sample interference levels by identifying fundamental frequencies within image frequency domains using discrete Fourier decomposition. The proposed approach identifies influential fundamental frequencies and perturbation amplitudes that affect sample attributes while minimizing disruptions. Injecting triggers into contaminated samples by utilizing these parameters can yield higher attack success rates without compromising accuracy for legitimate inputs. The method's efficacy is substantiated through exhaustive experimentation, surpassing prevailing techniques and demonstrating resilience against diverse defense strategies. The research exposes the susceptibility of deep learning models to backdoor attacks, providing insights into fortifying model defense and creating secure AI systems.

REFERENCES

[1] Z. Zhang et al., "Neurotoxin: Durable backdoors in federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 26429–26446.

[2] Y. Liu et al., "Defending label inference and backdoor attacks in vertical federated learning," 2021, *arXiv:2112.05409*.

[3] X. Gong, Y. Chen, H. Huang, Y. Liao, S. Wang, and Q. Wang, "Coordinated backdoor attacks against federated learning with model-dependent triggers," *IEEE Netw.*, vol. 36, no. 1, pp. 84–90, Jan./Feb. 2022.

[4] Y. Ge et al., "Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 826–834.

[5] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 11957–11965.

[6] Y. Chen, Z. Zheng, and X. Gong, "MARNet: Backdoor attacks against cooperative multi-agent reinforcement learning," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 5, pp. 4188–4198, Sep.–Oct. 2023.

[7] L. Wang, Z. Javed, X. Wu, W. Guo, X. Xing, and D. Song, "BACKDOORL: Backdoor attack against competitive reinforcement learning," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3699–3705.

[8] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *Proc. IEEE 7th Eur. Symp. Secur. Privacy*,, 2022, pp. 703–718.

[9] H. Phan, Y. Xie, J. Liu, Y. Chen, and B. Yuan, "Invisible and efficient backdoor attacks for compressed deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2022, pp. 96–100.

[10] X. Qi, T. Xie, R. Pan, J. Zhu, Y. Yang, and K. Bu, "Towards practical deployment-stage backdoor attack on deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13347–13357.

[11] J. Breier, X. Hou, M. Ochoa, and J. Solano, "FooBAR: Fault fooling backdoor attack on neural network training," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 3, pp. 1895–1908, May/Jun. 2023.

[12] Á. Berta, G. Danner, I. Hegedus, and M. Jelasity, "Hiding needles in a haystack: Towards constructing neural networks that evade verification," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, 2022, pp. 51–62.

[13] Y. Wang, E. Sarkar, W. Li, M. Maniatakos, and S. E. Jabari, "Stop-and-go: Exploring backdoor attacks on deep reinforcement learning-based traffic congestion control systems," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4772–4787, 2021.

[14] J. Chen, X. Wang, Y. Zhang, H. Zheng, S. Yu, and L. Bao, "Agent manipulator: Stealthy strategy attacks on deep reinforcement learning," *Appl. Intell.*, vol. 53, no. 10, pp. 12831–12858, 2023.

[15] Y. Feng, B. Ma, J. Zhang, S. Zhao, Y. Xia, and D. Tao, "FIBA: Frequency-injection based backdoor attack in medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20876–20885.

[16] T. Wang, Y. Yao, F. Xu, S. An, H. Tong, and T. Wang, "An invisible black-box backdoor attack through frequency domain," in *Proc. 17th Eur. Conf. Comput. Vis.*,, 2022, pp. 396–413.

[17] R. Hou, T. Huang, H. Yan, L. Ke, and W. Tang, "A stealthy and robust backdoor attack via frequency domain transform," *World Wide Web*, vol. 26, pp. 1–17, 2023.

[18] H. A. A. K. Hammoud and B. Ghanem, "Check your other door! Creating backdoor attacks in the frequency domain," 2021, *arXiv:2109.05507*.

[19] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, vol. 1. Princeton, NJ, USA: Princeton Univ. Press, 2011.

[20] J. W. Goodman, *Introduction to Fourier Optics*. Greenwood Village, CO, USA: Roberts & Co., 2005.

[21] R. Wightman, H. Touvron, and H. Jégou, "ResNet strikes back: An improved training procedure in timm," in *Proc. NeurIPS 2021 Workshop ImageNet: Past, Present, Future*, 2021.

[22] S. Zhao, L. Zhou, W. Wang, D. Cai, T. L. Lam, and Y. Xu, "Toward better accuracy-efficiency trade-offs: Divide and co-training," *IEEE Trans. Image Process.*, vol. 31, pp. 5869–5880, 2022.

[23] M. D. McDonnell and T. Vladusich, "Enhanced image classification with a fast-learning shallow convolutional neural network," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2015, pp. 1–7.

[24] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Master's thesis, Univ. Tront, 2009.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[26] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, 2008.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

[29] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526*.

[30] B. Wang et al., "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2019, pp. 707–723.

[31] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, 2019, pp. 113–125.

[32] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.