Harmless Copyright Protection for CT Dataset

Fengzhi Xu[1], Ziyuan Yang[1], Yan Liang[1], Yingyu Chen[2], Wenjun Xia[3], and Yi Zhang[1]*

[1] School of Cyber Science and Engineering, Sichuan University, Chengdu, China
[2] College of Computer Science, Sichuan University, Chengdu, China
[3] School of Computer Science and Engineering, Southeast University, Nanjing, China

Abstract. Dataset is essential for advancing deep learning in healthcare, yet its unauthorized use raises serious concerns regarding patient privacy leakage and intellectual property infringement. Existing copyright protection methods typically rely on backdoor-triggered verification mechanisms, but these methods inherently introduce security vulnerabilities by embedding malicious backdoor behaviors that can lead to abnormal outputs. To address this issue, we propose a novel harmless copyright verification threat model. Unlike traditional methods that depend on trigger-activated abnormal outputs as verification signals, our approach instead uses prediction accuracy on watermarked samples as the verification criterion. The core idea is to enforce improved performance on trigger-injected samples, enabling reliable ownership verification without inducing malicious behaviors. Specifically, we propose a Learnable Measurement-Domain Harmless Backdoor Watermark (LMHBW) method. Unlike traditional image-domain watermarking techniques, which operate on reconstructed images that are easily accessible and thus vulnerable to tampering, we shift the protection to the measurement domain prior to image reconstruction. Motivated by the fact that raw measurement data are typically confined within imaging devices (e.g., scanners) and are rarely exposed externally, this design enhances the stealthiness and robustness of the embedded watermark against detection and removal. Experimental results demonstrate that the proposed threat model enables effective and reliable watermark verification while remaining behaviorally harmless, and further validate the feasibility and advantages of measurement-domain watermarking for medical imaging copyright protection.

Keywords: Copyright Protection · Backdoor Watermark · Computed Tomography · Measurement-domain Watermark · Harmless Verification.

1 Introduction

Datasets have gained widespread attention and usage with the rapid advancement of deep learning in healthcare [14]. However, medical datasets often contain sensitive information about individuals, and their unauthorized use may lead to

* Corresponding author: Yi Zhang (yzhang@scu.edu.cn)

serious privacy leakage concerns [16]. To mitigate the misuse of medical datasets, it is essential to develop effective copyright protection techniques.

Existing copyright protection methods can be broadly classified into three categories. 1) Encryption-based methods [12, 11], which secure datasets or sensitive information prior to release, but they often significantly restrict data accessibility and downstream usability. 2) Traditional digital watermarking techniques [3, 4], which embed identifiable signals into data for subsequent ownership verification, but are generally limited to verifying dataset ownership and cannot determine whether a downstream model has been trained on the protected data. 3) Backdoor watermarking, which address this limitation by injecting backdoor triggers into training data. Models trained on such data inherit the backdoor behavior, enabling ownership verification by activating the trigger during inference on suspicious models.

Backdoor watermarking has attracted significant attention, and several representative approaches have been proposed. Adi *et al.* [1] pioneered the use of backdoors as a tool for model ownership verification by embedding watermarks via randomly mislabeled trigger samples during training. Chattopadhyay *et al.* [6] uses adversarial trigger samples and distributes watermark information across all network layers to improve robustness. Bansal *et al.* [5] proposed a certifiable trigger-based watermark that leverages randomized smoothing to theoretically guarantee robustness against parameter perturbations within an l_2-norm ball. In addition, Untargeted Backdoor Watermark (UBW) introduces an untargeted backdoor watermark that reduces the reliance on deterministic target labels by increasing prediction variability on watermarked samples [10]. Despite their effectiveness, these methods typically rely on backdoor-triggered verification mechanisms, where a specific trigger is embedded into the input space to force the model to produce a predefined abnormal output, which serves as the verification signal. While effective for ownership verification, this paradigm inherently introduces malicious backdoor behavior, as it enables intentional manipulation of model predictions under trigger activation, thereby raising potential security risks. In particular, adversaries may exploit such backdoors to induce unintended or harmful outputs, undermining the trustworthiness of the deployed model.

To address this limitation, we propose a novel harmless copyright verification threat model. Different from traditional methods that rely on abnormal outputs activated by backdoor triggers, our approach uses the prediction accuracy on watermarked samples as the verification signal. The core idea is to explicitly enforce improved prediction performance on trigger-injected samples, rather than inducing incorrect or abnormal outputs. In this way, ownership can be reliably verified without embedding malicious backdoor behavior into the model, as shown in Fig.1.

Specifically, we propose a Learnable Measurement-Domain Harmless Backdoor Watermark (LMHBW) for copyright protection of CT datasets. CT data exists in two representations: the pre-imaging measurement-domain and the post-imaging image domain. Measurement data are acquired through the physical scanning process and are typically confined within the imaging device, whereas

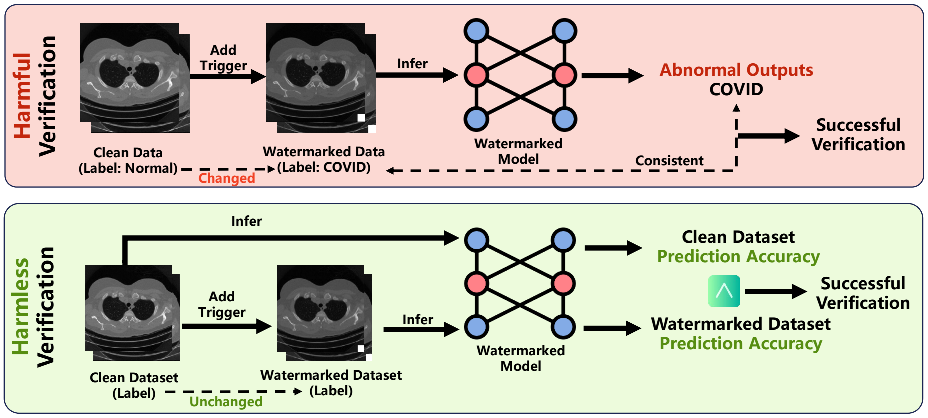

Fig. 1. Comparison between the proposed harmless verification threat model and existing harmful verification models.

reconstructed images are routinely exported for clinical diagnosis and thus highly accessible. Unlike traditional image-domain protection methods, which are vulnerable to inspection, tampering, or removal, measurement-domain mechanisms are inherently more stealthy and significantly harder to detect and manipulate. Therefore, we focus on embedding watermarks in the measurement domain.

In clinical practice, diagnosis is performed on reconstructed images that clearly reveal anatomical structures, rather than on raw measurement data. Since measurement data must undergo a complex inverse reconstruction process before being usable, designing effective measurement-domain perturbations without degrading reconstruction quality is non-trivial [18]. To address this challenge, we introduce a learnable measurement-domain watermark. Our approach is guided by two objectives. First, we enforce reconstruction consistency, ensuring that images reconstructed from watermarked measurements remain visually indistinguishable from those obtained from clean measurements. Second, we ensure downstream verifiability, such that the embedded watermark can be reliably identified through prediction accuracy on watermarked samples. To achieve this, we design a dedicated loss function and optimize the measurement-domain watermark using a surrogate reconstruction model, jointly balancing image fidelity and verification effectiveness. After optimization, the learned watermark is injected into a subset of clean training data, forming a protected dataset. In downstream learning tasks, models trained on this dataset exhibit higher prediction accuracy on watermarked samples compared to clean samples, enabling reliable ownership verification. Our contributions are summarized as follows:

– We propose a novel harmless copyright verification threat model, which replaces abnormal backdoor-triggered outputs with prediction accuracy on watermarked samples as the verification signal, enabling ownership verification without introducing malicious model behavior.

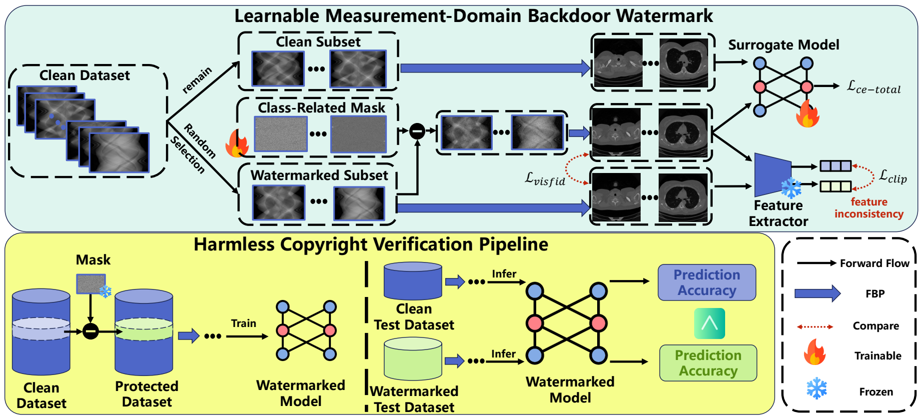

Fig. 2. Overview of the proposed framework.

- We introduce LMHBW, a learnable measurement-domain watermarking method that optimizes a mask in the CT measurement space, preserving reconstruction quality while enabling effective downstream verification.
- Extensive experiments demonstrate the effectiveness of the proposed harmless verification paradigm and validate the superior stealthiness and practicality of measurement-domain watermarking.

2 Methodology

2.1 Overview

To achieve harmless copyright protection, we adopt prediction accuracy on watermarked samples as the verification signal, instead of abnormal outputs induced by backdoor triggers. We propose a Learnable Measurement-Domain Harmless Backdoor Watermark (LMHBW) framework, as illustrated in Fig. 2, which consists of two stages: watermark optimization and verification. Section 2.2 describes the measurement-domain watermark optimization. The optimization is formulated with two objectives: (i) preserving reconstruction consistency between watermarked and clean measurements, and (ii) ensuring verifiability via downstream prediction accuracy on watermarked samples. Section 2.3 introduces the verification pipeline, where the learned watermark is embedded into the measurement data and ownership is verified through accuracy differences on watermarked versus clean samples.

2.2 Measurement-Domain Backdoor Watermark

Since measurement data remains within the scanner and is rarely exposed, we focus on pre-imaging measurement-domain watermarking, which is inherently resistant to detection and removal. However, measurement data must be reconstructed into images for clinical diagnosis. Due to the complexity of medical image reconstruction, manually designed measurement-domain perturbations may degrade reconstructed image quality and thus affect diagnosis. To mitigate this issue, we introduce a learnable class-related watermark mask $m \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of classes.

The watermark is optimized on the clean dataset, which is randomly split into a watermarked subset S and a clean subset R according to a watermarking ratio α. Samples $s \in S$ are embedded with class-specific watermarks for copyright protection. The optimization is guided by two objectives: (i) preserving visual consistency of reconstructed images to avoid diagnostic interference, and (ii) enabling reliable verification via prediction accuracy on watermarked samples.

To preserve visual consistency, we define:

$$\mathcal{L}_{visfid} = \sum_{s\in S}(w_1\mathcal{L}_{mse}(\mathcal{T}(s), \mathcal{T}(s-m_{c_s}))+w_2(1-\mathcal{L}_{ssim}(\mathcal{T}(s), \mathcal{T}(s-m_{c_s})))), \quad (1)$$

where w_1 and w_2 are weighing parameters, $\mathcal{T}$ denotes the reconstruction operator (FBP in this work), m_{c_s} is the class-specific watermark and c_s is the lable of sample s. The first term enforces pixel-level consistency, while the second preserves structural similarity [19].

To ensure downstream verifiability, we aim to induce a discriminative gap between watermarked and clean samples. Specifically, we introduce feature perturbation using a pretrained CLIP encoder Φ [13]:

$$\mathcal{L}_{clip} = w_3 \sum_{s\in S} \mathrm{CosSim}(\Phi(\mathcal{T}(s)), \Phi(\mathcal{T}(s - m_{c_s}))), \quad (2)$$

which encourages distinguishable feature representations between clean and watermarked reconstructions. w_3 is the weight for $\mathcal{L}_{clip}$. Finally, to ensure correct classification behavior while preserving clean accuracy, we train a surrogate model F with:

$$\mathcal{L}_{ce-total} = w_4 \sum_{s\in S} \mathcal{L}_{ce}(F(\mathcal{T}(s - m_{c_s})), c_s) + w_4 \sum_{r\in R} \mathcal{L}_{ce}(F(\mathcal{T}(r)), c_r), \quad (3)$$

where w_4 is the weight for $\mathcal{L}_{ce}$, $r \in R$ denotes clean samples and c_r denotes the lable of sample r.

Overall, the optimization objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{visfid} + \mathcal{L}_{clip} + \mathcal{L}_{ce-total}. \quad (4)$$

2.3 Harmless Copyright Verification Pipeline

After watermark optimization, we construct a harmless copyright verification pipeline based on the proposed threat model. The pipeline consists of two stages: watermark embedding during training and watermark verification during inference.

Watermark Embedding. A subset S of the clean training dataset is selected and embedded with the optimized watermark, which is then combined with the clean subset R to form the protected dataset P:

$$P = R \cup \{s - m_{c_s} | s \in S\}. \quad (5)$$

Models trained on P inherit the watermark during learning.

Watermark Verification. During inference, we construct a clean test set and a corresponding watermarked test set, where the latter is generated by applying the same watermark to the clean test samples. Copyright ownership is verified by comparing prediction accuracy: if a suspicious model achieves higher accuracy on the watermarked test set than on the clean test set, we infer that it was trained on the protected dataset P; otherwise, ownership is not confirmed. This accuracy discrepancy serves as the verification signal for dataset copyright protection.

3 Experiment

3.1 Experiment Settings

Dataset. The proposed method is evaluated on the public COVID-19 Low-Dose CT dataset [2] collected from Babak Imaging Center, Iran. The dataset is organized at the subject level, where each subject contains multiple CT slices. The clean training dataset includes 15 COVID-19 subjects and 15 normal subjects, while the clean test set consists of 5 COVID-19 subjects and 5 normal subjects. This subject-level split ensures that training and test data are separated by patients, avoiding data leakage across sets.

Evaluation Metrics. Benign Accuracy (BA) is defined as the classification accuracy of the model on the clean test set, reflecting its performance on unmodified data. Verification Success Rate (VSR) evaluates the effectiveness of ownership verification. We conduct 20 independent runs, each consisting of training, testing, and verification. VSR is defined as the proportion of correctly verified watermarked models among all ground-truth watermarked models. Watermark Success Rate (WSR) measures the classification accuracy of the model on the watermarked test set. To quantitatively assess the harmfulness of different backdoor watermarking methods, we adopt the prediction consistency indices C_b and C_w from [15]. Specifically, C_b measures the consistency of predictions between clean and watermarked models on clean samples, while C_w measures the same consistency on watermark samples.

Implementation Details. All neural networks are implemented in PyTorch and trained on an NVIDIA RTX 4090 GPU. We adopt the widely used FBP as the reconstruction operator $\mathcal{T}$, implemented using CTLib [17]. For fair comparison across different methods, we consistently used ResNet-18 [9] as the backbone classifier in downstream tasks. The watermarking ratio is set to $\alpha = 0.2$ in all experiments.

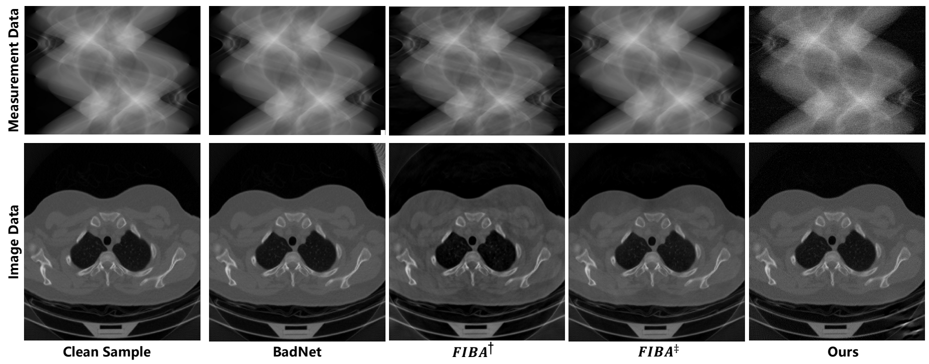

Fig. 3. Original and watermarked measurement data and their corresponding reconstructed images.

Table 1. Quantitative comparison of different methods. Harmful results ($C_w < 20\%$) or low verification success rate ($VSR < 70\%$) are highlighted in red.

Methods	BA(%)↑	VSR(%)↑	C_b(%)↑	C_w(%)↑
Clean	68.42±2.88	-	-	-
BadNet	69.62±4.81	60	93.21±5.54	6.35±4.98
FIBA†	68.79±6.69	60	90.51±6.84	8.83±4.69
FIBA‡	67.17±2.97	65	94.39±3.66	11.04±5.43
Ours	69.30±5.40	100	92.61±4.05	58.15±20.05

3.2 Results

To validate the effectiveness and harmlessness of the proposed method, we conduct comparative experiments against representative baselines. In our experiments, "Clean" denotes a baseline model trained solely on clean data. The proposed method is compared with backdoor-based watermarking methods that rely on trigger-induced verification, where trigger patterns are designed following representative backdoor attacks such as BadNet [8] and FIBA [7]. Specifically, "FIBA†" and "FIBA‡" denote variants using triggers derived from the amplitude spectrum of reconstructed images and measurement data, respectively. As shown in Tab. 1, the proposed method achieves high VSR, demonstrating effective ownership verification. Meanwhile, it attains the highest C_w, which indicates strong prediction consistency between clean and watermarked models on watermark samples, while avoiding abnormal output behavior.

To further evaluate visual fidelity in the image domain, we compare the effects of different backdoor triggers on reconstructed images. As illustrated in Fig. 3, BadNet introduces localized perturbations in the top-right region, while FIBA† induces global distortions. In contrast, both FIBA‡ and our method preserve structures closer to the original images, demonstrating superior visual consistency and reduced perceptual degradation.

Table 2. Quantitative results across different networks.

	ResNet		EfficientNet		MobileNet	
Surrogate model	BA(%)↑	VSR(%)↑	BA(%)↑	VSR(%)↑	BA(%)↑	VSR(%)↑
ResNet	69.30±5.40	100	73.73±4.76	100	78.72±6.37	100
EfficientNet	68.90±4.77	90	73.03±4.75	100	79.40±4.45	100
MobileNet	68.73±4.88	95	70.88±4.45	100	79.62±6.02	100

Table 3. Quantitative results with different watermarking ratios (α) using ResNet.

α	BA(%)↑	WSR(%)↑	VSR(%)↑
0.05	68.76±4.65	70.62±5.34	60
0.1	69.06±4.62	87.57±7.23	100
0.2	69.30±5.40	89.90±6.14	100
0.3	69.78±4.14	76.02±3.91	95
0.4	68.16±3.57	82.96±5.49	100

To evaluate the generalization of the proposed method across network architectures, we conduct cross-architecture experiments. As shown in Tab. 2, the rows represent surrogate models used for watermark optimization, while the columns denote testing models used for verification. The proposed method maintains high VSR even when the surrogate and testing architectures differ, demonstrating strong cross-architecture generalization.

We further analyze the effect of the watermarking ratio α to examine whether reliable verification can be achieved by modifying only a small portion of the dataset. Tab. 3 reports the results under different watermarking ratios. Even when only 10% of the training data are watermarked, the proposed method achieves a high verification success rate. When $\alpha > 10\%$, the VSR remains consistently above 95%. This indicates that the proposed method enables effective dataset copyright verification with only limited data modification.

4 Conclusion

In this paper, we present a novel harmless copyright verification threat model. Unlike conventional backdoor-based verification methods that rely on trigger-induced abnormal predictions, our framework eliminates the need for malicious output behaviors and instead leverages prediction accuracy as the verification signal, thereby mitigating security risks associated with backdoor activation. Specifically, we introduce a Learnable Measurement-Domain Harmless Backdoor Watermark (LMHBW) that embeds watermarks in the measurement domain prior to image reconstruction. This design improves stealthiness while enabling reliable ownership verification through prediction accuracy in downstream tasks. Experimental results demonstrate that the proposed threat model achieves effective watermark verification while remaining behaviorally harmless. Moreover, extensive evaluations further validate the effectiveness and generalization capability of measurement-domain watermarking.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under 62271335, in part by the Sichuan Science and Technology Program under Grant 2025ZNSFSC0470.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adi, Y., Baum, C., Cisse, M., Pinkas, B., Keshet, J.: Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In: 27th USENIX security symposium (USENIX Security 18). pp. 1615–1631 (2018)
2. Afshar, P., Rafiee, M.J., Naderkhani, F., Heidarian, S., Enshaei, N., Oikonomou, A., Babaki Fard, F., Anconina, R., Farahani, K., Plataniotis, K.N., et al.: Human-level covid-19 diagnosis from low-dose ct scans using a two-stage time-distributed capsule network. Scientific Reports **12**(1), 4827 (2022)
3. Asnani, V., Collomosse, J., Bui, T., Liu, X., Agarwal, S.: Promark: Proactive diffusion watermarking for causal attribution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10802–10811 (2024)
4. Asnani, V., Kumar, A., You, S., Liu, X.: Probed: Proactive object detection wrapper. Advances in Neural Information Processing Systems **36**, 77993–78005 (2023)
5. Bansal, A., Chiang, P.y., Curry, M.J., Jain, R., Wigington, C., Manjunatha, V., Dickerson, J.P., Goldstein, T.: Certified neural network watermarks with randomized smoothing. In: International Conference on Machine Learning. pp. 1450–1465. PMLR (2022)
6. Chattopadhyay, N., Chattopadhyay, A.: Rowback: Robust watermarking for neural networks using backdoors. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1728–1735. IEEE (2021)
7. Feng, Y., Ma, B., Zhang, J., Zhao, S., Xia, Y., Tao, D.: Fiba: Frequency-injection based backdoor attack in medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20876–20885 (2022)
8. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. Ieee Access **7**, 47230–47244 (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Li, Y., Bai, Y., Jiang, Y., Yang, Y., Xia, S.T., Li, B.: Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. Advances in Neural Information Processing Systems **35**, 13238–13250 (2022)
11. Li, Y., Liu, P., Jiang, Y., Xia, S.T.: Visual privacy protection via mapping distortion. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3740–3744. IEEE (2021)
12. Martins, P., Sousa, L., Mariano, A.: A survey on fully homomorphic encryption: An engineering perspective. ACM Computing Surveys (CSUR) **50**(6), 1–33 (2017)

13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
14. Serte, S., Serener, A., Al-Turjman, F.: Deep learning in medical imaging: A brief review. Transactions on Emerging Telecommunications Technologies **33**(10), e4080 (2022)
15. Sun, M., Wang, R., Zhu, Z., Jing, L., Guo, Y.: Entropymark: Towards more harmless backdoor watermark via entropy-based constraint for open-source dataset copyright protection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30692–30701 (2025)
16. Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., Wang, G.: Security and privacy in the medical internet of things: a review. Security and Communication Networks **2018**(1), 5978636 (2018)
17. Xia, W., Lu, Z., Huang, Y., Shi, Z., Liu, Y., Chen, H., Chen, Y., Zhou, J., Zhang, Y.: Magic: Manifold and graph integrative convolutional network for low-dose ct reconstruction. IEEE Transactions on Medical Imaging (2021)
18. Yang, Z., Chen, Y., Sun, M., Zhang, Y.: Inject backdoor in measured data to jeopardize full-stack medical image analysis system. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 393–402. Springer (2024)
19. You, C., Yang, Q., Shan, H., Gjesteby, L., Li, G., Ju, S., Zhang, Z., Zhao, Z., Zhang, Y., Cong, W., et al.: Structurally-sensitive multi-scale deep neural network for low-dose ct denoising. IEEE access **6**, 41839–41855 (2018)