Adversarial2Adversarial: Defending against Adversarial Fingerprint Attacks without Clean Images

Pyo Min Hong, So Hyun Kang, Jinhyeon Kim, Ji Hoo Kim, and Youn Kyu Lee*
Department of Computer Engineering
Hongik University
Seoul, Republic of Korea
pyomindl@g.hongik.ac.kr, th_gus4734@g.hongik.ac.kr, tbutc@g.hongik.ac.kr,
jhkim990212@g.hongik.ac.kr, younkyul@hongik.ac.kr

***Abstract*—A number of denoising-based methods have been proposed to defend against adversarial fingerprint attacks. However, these methods inherently rely on having a clean image that corresponds to each adversarial fingerprint image. In this paper, we propose a novel denoising-based defense method without the need for clean fingerprint images. Our approach leverages a Noise2Noise mechanism, which performs denoising based on the noisy dataset. This enables us to effectively eliminate any adversarial noise that may be embedded in fingerprint images without training on clean fingerprint images. The experimental results on real-world datasets confirm that our method is robust against untrained adversarial fingerprint attacks while outperforming existing methods.**

***Index Terms*—fingerprint liveness detection, adversarial attack, deep learning, denoising**

I. Introduction

With the continuous advancement of deep learning technology in fingerprint liveness detection, fingerprint authentication systems are increasingly being employed in various domains, including user authentication [1] and access control [2] [3]. However, recently, there have been cases of adversarial fingerprint examples observed, where adversarial noises intentionally induce misclassification of target deep learning models for fingerprint liveness detection. These adversarial fingerprint attacks pose significant security concerns, such as privacy breaches, identity theft, and financial losses [4] [5].

To address these issues, various methods have been proposed, such as detecting adversarial examples and isolating them from the input data, as well as generating models trained on adversarial examples [6]. Moreover, denoising-based approaches have been proposed, which rely on the elimination of adversarial noise through an image reconstruction mechanism [7] [8]. However, existing denoising-based methods typically require labeled data in the form of pairs of noisy images and clean images, which can be time-consuming and costly to acquire. Additionally, they may exhibit degraded performance for untrained types of adversarial noises [9]. Consequently, these existing methods do not provide robust defense performance against adversarial fingerprint attacks.

*Corresponding Author

In this paper, we propose a novel denoising-based defense method that leverages Noise2Noise (N2N), a denoising technique that can be trained without the need for clean images. The N2N mechanism [10] involves the random generation of two noisy images, each following a zero-mean distribution (not necessarily the same) of an original image. These generated images are then utilized as the source and target for reconstructing the original image. In other words, N2N performs denoising by training a mapping from a noisy image (i.e., source image) to the other noisy image (i.e., target image). Existing denoising-based methods were limited in their ability to effectively defend against various types of adversarial attacks when the required corresponding clean images for training were unavailable due to insufficient adversarial example datasets. However, our approach overcomes these limitations by leveraging N2N, which does not rely on clean images. Based on the N2N mechanism, our proposed method effectively eliminates any adversarial noise embedded in fingerprint images by leveraging the information learned from adversarial examples. Subsequently, the denoised fingerprint images are passed to the classifier, enabling a robust liveness detection against adversarial attacks.

The paper makes the following contributions: (1) proposing a novel denoising-based defense method that eliminates the reliance on clean images; (2) achieving robust defense performance against a range of adversarial attacks; (3) implementing a prototype of our proposed method; and (4) validating the effectiveness of our method using a real-world dataset.

The organization of this paper is as follows: Section 2 discusses the related work, Section 3 presents our proposed method, Section 4 describes our evaluation, and Section 5 provides the conclusions.

II. Related Work

A. Adversarial attack

An adversarial attack is a specific type of attack that manipulates input images by adding adversarial noise. The adversarial noise leads the target model to misclassify as the attacker intends, and it is imperceptible to the naked

eye. Goodfellow et al. [11] proposed the Fast Gradient Sign Method (FGSM), which entails identifying the gradient of the loss function for an input image and generating noise in the direction that increases the loss. Madry et al. [12] introduced Projected Gradient Descent (PGD), which involves updating the adversarial noise by performing FGSM in multiple steps.

B. Defend against adversarial attacks with denoising

Denoising refers to the process of eliminating or reducing noise in an image while preserving its important features. Denoising is being widely used to recover the quality of videos or images [13], or to defend against adversarial attacks by eliminating adversarial noise [14]. F. Liao et al. [14] proposed High-Level Representation Guided Denoiser (HGD), which employs a denoiser trained with a loss function that quantifies the difference between the original and adversarial examples. T. Dai et al. [15] introduced Deep Image Prior Driven Defense (DIPDefend), which reconstructs images without prior learning by leveraging the properties of deep image prior. This approach successfully eliminates adversarial noise while accurately preserving the intrinsic structures of clean images. Y. Bakhti et al. [16] proposed Deep Denoising Sparse Autoencoder (DDSA), which eliminates or reduces adversarial noise by employing dimensionality reduction in the image preprocessing step. Denoising is widely employed as a defense mechanism against adversarial attacks, but a reliable defense method for adversarial fingerprint attacks has not yet been proposed. Moreover, to train denoising models capable of addressing a range of adversarial fingerprint attacks, it is essential to obtain clean images corresponding to each attack. However, obtaining clean images from adversarial fingerprint images poses a considerable challenge due to the diverse types of adversarial attacks, making the process costly. Hence, it is required to design a new method that can effectively protect against adversarial fingerprint attacks by training solely on adversarial fingerprint images, without requiring clean images.

III. Our Approach

In this paper, we propose a novel denoising-based defense method against adversarial fingerprint attacks. Our proposed method utilizes Noise2Noise (N2N), which effectively eliminates the adversarial noise from the given fingerprint images by performing denoising based on the noisy dataset. Our proposed method provides robust defense performance against adversarial fingerprint attacks without requiring clean images for each attack image. As shown in Fig. 1, our proposed method consists of two distinct steps. (1) N2N-based defense model generation: A defense model is generated by training on adversarial fingerprint datasets that contain various adversarial noise additions, enabling effective denoising of the adversarial noise. (2) N2N-based fingerprint liveness detection: Given a fingerprint image, the defense model performs denoising, and subsequently, a classifier determines the liveness of the denoised image.

A. N2N-based defense model generation

In this step, a N2N-based defense model is trained using a dataset composed of pairs of adversarial fingerprint images with different adversarial noise additions. Specifically, in the training phase, adversarial fingerprint images that include the fingerprint region along with its surrounding margin are used. During the acquisition of fingerprint images, factors such as sensor types, fingerprint size, and acquisition methods can lead to the inclusion of margins outside the fingerprint region in the acquired images. These margin areas can serve as significant embedding targets for adversarial noise. For each pair in the dataset, which includes the source and target images as illustrated in Fig. 1, our method involves cropping the margins of both images and utilizing them as a pair to train the N2N-based defense model.

$$\arg\min_{\theta} \sum_{i} L2(f_\theta(\hat{x}_i), \hat{y}_i), \tag{1}$$

where $\hat{x}_i$ represents the source margin image, $\hat{y}_i$ represents the target margin image, and f_θ represents a collection of parametric mappings. The N2N-based defense model is trained by minimizing the difference between the images using a loss function, $L2$, which calculates the squared difference between the images and then computes the average across all pairs in the dataset. In the inference phase, the trained model is utilized to eliminate adversarial noise from the given fingerprint image.

B. N2N-based fingerprint liveness detection

In this step, the trained N2N-based defense model is utilized to eliminate adversarial noise, followed by performing liveness detection. The N2N-based defense model reconstructs a given fingerprint image, which is then examined by a liveness detection classifier. This classifier is trained on original fingerprints without embedded noise and infers a liveness score for each given fingerprint image. The inferred score is subsequently compared with a pre-defined threshold to determine the liveness status (i.e., alive or fake).

IV. Experimental Evaluation

To evaluate the effectiveness of our proposed method, we conducted evaluations based on the following research questions:

- **RQ#1**: Does our proposed method provide effective defense performance against "untrained" adversarial fingerprint attacks?
- **RQ#2**: Does our proposed method provide superior performance compared to the existing denoising methods?

A. Experimental Setting

To evaluate our method, we selected the LivDet 2015 dataset [17], a real-world dataset that includes various fingerprint images acquired from different types of sensors (i.e., Green Bit, Digital Persona, Biometrika, and Crossmatch). Among those, we selected the Green Bit dataset due to its high vulnerability to adversarial attacks, including *FGSM* and *PGD* [18]. The Green Bit dataset consists of 2,000 training

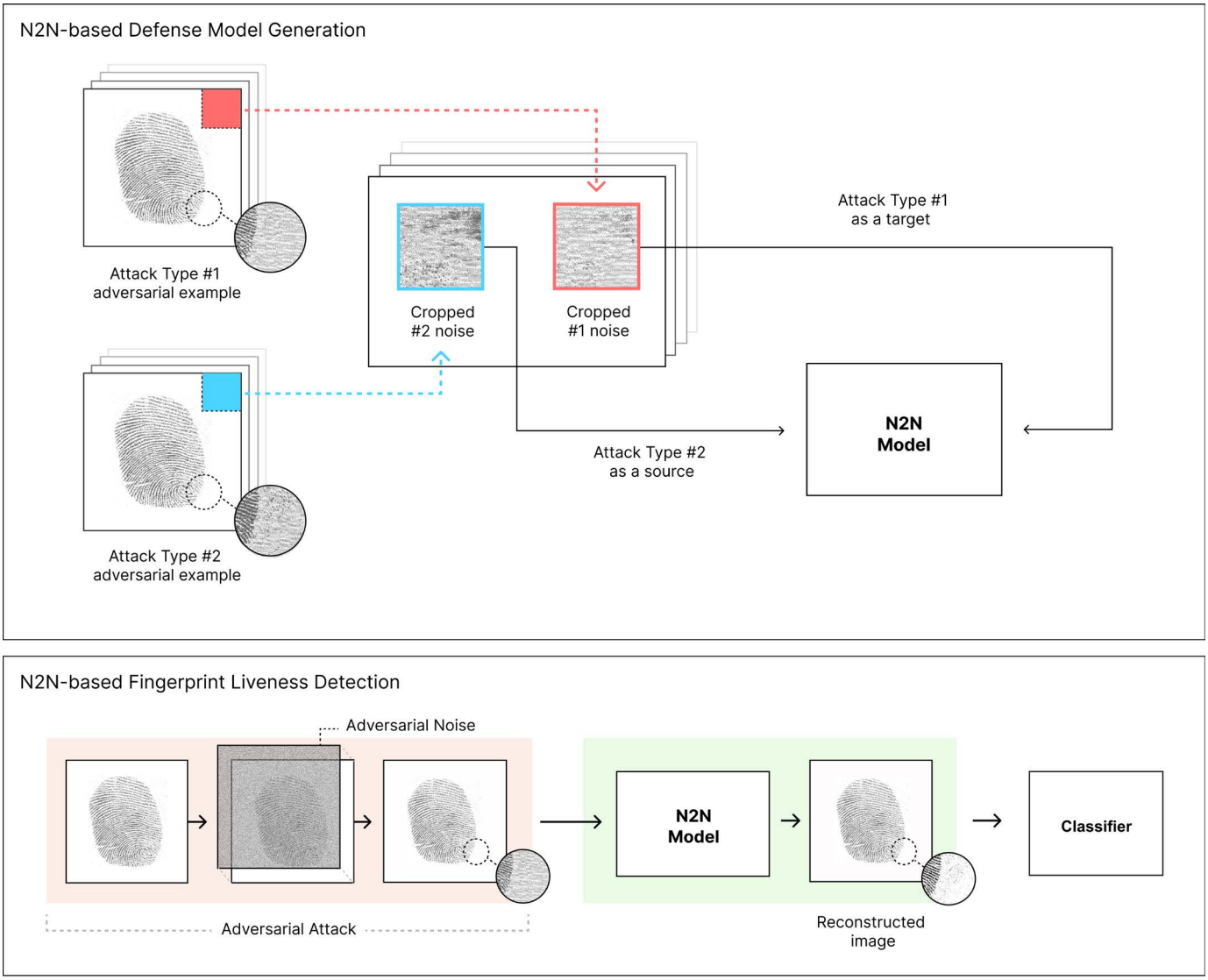

Fig. 1. An Overview of Our Proposed Method

images (Alive: 1,000, Fake: 1,000) and 2,500 testing images (Alive: 1,000, Fake: 1,500), each with a size of 500×500 pixels. For our target liveness detection classifier, we employed ResNet-50, which has demonstrated superior performance in fingerprint liveness detection [19]. With the ResNet-50 model trained on the training images from the Green Bit dataset, we compared its liveness detection performance on the original testing images, testing images with adversarial attacks, and testing images denoised by our methods.

To train our N2N-based defense model, we utilized a total of 2,000 images from the Green Bit dataset, consisting of 800 images from the "alive" class and 1,200 images from the "fake" class. For each fingerprint image, we created two adversarial images using *PGD* (eps=0.3, alpha=2/255, iters=40) and *FGSM* (eps=0.03) methods, respectively. Subsequently, we cropped the non-fingerprint areas of each adversarial image to a size of 128 × 128 and used them as training pairs. The hyperparameters for the adversarial attacks were defined by the authors.

We validated RQ#1 to evaluate the robustness of our proposed method against untrained adversarial fingerprint attacks. We assessed its defensive performance on adversarial images that were not used during the training process. To generate untrained adversarial fingerprint images, we adjusted the hyperparameters for each adversarial attack as follows: *PGD* (eps=0.4, alpha=2/255, iters=40) and *FGSM* (eps=0.04).

We validated RQ#2 to assess the effectiveness of our proposed method. We conducted a comparative analysis with existing denoising methods, Deep Image Prior (*DIP*) [20] and Self2Self (*S2S*) [21]. For *DIP* and *S2S*, we used the hyperparameters as defined by the respective authors: *DIP* (learning rate=0.01, loss function=MSE, and iteration=800) and *S2S* (learning rate=0.0001, prediction=100, and step=1000). For each adversarial attack (i.e., *PGD* and *FGSM*), we employed the same hyperparameters as RQ#1: *PGD* (eps=0.4, alpha=2/255, iters=40) and *FGSM* (eps=0.04).

TABLE I
THE RESULT OF OUR EXPERIMENT

Baseline	PGD [12]		FGSM [11]	
	Adversarial Attack	Our Method	Adversarial Attack	Our Method
93.16	0.00	**64.99**	60.00	**79.79**

The models used in the experiment were trained in the following environments—N2N and *DIP*: NVIDIA GeForce RTX 3090, python 3.8.0, pytorch 1.11.0+cu113; *S2S* and Resnet-50: NVIDIA GeForce RTX 3090, python 3.8.0, Tensorflow 2.7.0. The hyperparameters used in the experiments were set based on the parameters defined by the authors—N2N: Adam optimizer, 0.001 learning rate, and 50 epochs; *DIP*: MSE loss function, 0.01 learning rate, and 800 iterations; *S2S*: 100 prediction, 0.0001 learning rate, and 1000 steps; ResNet-50: Adam optimizer, 0.001 learning rate, 8 batch size and 20 epochs.

B. Experimental Results

(RQ#1) Evaluation of the robustness towards untrained adversarial fingerprint attacks: To validate RQ#1, we implemented each adversarial attack (i.e., *PGD* and *FGSM*) using hyperparameters that were not used during the training process. Table I provides a detailed presentation of the results. The baseline performance, which refers to the liveness detection performance on the clean image, was 93.16%. However, when applying an adversarial fingerprint attack using *PGD*, the baseline performance decreased to 0.00%. On the other hand, by applying our proposed method to the adversarial images, the performance increased to 64.99%. Similarly, when performing an adversarial fingerprint attack using *FGSM*, the baseline performance decreased to 60.00%. However, by applying our proposed method to the adversarial images, the performance increased to 79.79%.

The liveness detection performance of the adversarial images using *PGD* increased by 64.99%p when applying our proposed method. In the case of adversarial images using *FGSM*, the performance increased by 19.79%p when applying our proposed method. Although the increase in the liveness detection performance is relatively low in the case of *FGSM*, this is primarily because the liveness detection performance on the adversarial images was already relatively high at 60.00%. However, in the case of *PGD*, the improvement in liveness detection performance was significant. These results confirm the robustness of our proposed method against untrained adversarial fingerprint attacks.

(RQ#2) Comparison of the defense performance against adversarial fingerprint attacks between our method and existing denoising methods: To validate RQ#2, we conducted a comparative analysis with existing denoising methods, *DIP* and *S2S*. We employed each adversarial fingerprint attack (i.e., *PGD* and *FGSM*), using hyperparameters that were not used during the training process. Fig. 2 provides a detailed presentation of the results. The liveness detection performance

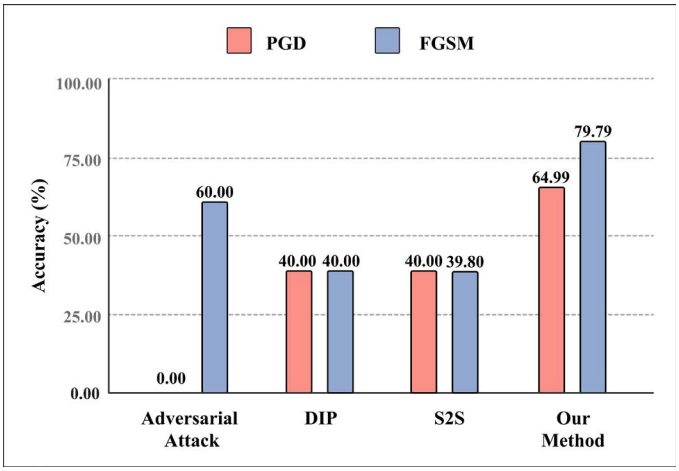

Fig. 2. A performance comparison of presentation attack detection between existing image reconstruction methods (DIP [20], S2S [21]) and our proposed method for each adversarial attack (PGD [12], FGSM [11]).

on adversarial images without denoising was 0.00% for *PGD* and 60.00% for *FGSM*, respectively. When reconstructing the adversarial images using *DIP*, the liveness detection performance was 40.00% for *PGD* and 40.00% for *FGSM*, respectively. When reconstructing the adversarial images using *S2S*, the liveness detection performance was 40.00% for *PGD* and 39.80% for *FGSM*, respectively. However, when reconstructing the adversarial images using our proposed method, the liveness detection performance significantly improved to 64.99% for *PGD* and 79.79% for *FGSM*, respectively.

V. CONCLUSION

In this paper, we proposed a new defense method that effectively eliminates adversarial noises added to fingerprint images by training solely on adversarial fingerprint images without relying on clean images. To achieve this, we employed N2N as the basis of our approach. Our experimental results confirm that our method is robust against untrained adversarial fingerprint attacks. Furthermore, our method significantly improves the compromised liveness detection performance, which is degraded by adversarial fingerprint attacks, in comparison to existing denoising methods (i.e., *DIP* and *S2S*). Our future work involves conducting extended evaluations on various types of adversarial attacks and fingerprint datasets. Additionally, we plan to assess the applicability of the proposed method to adversarial attacks in face recognition systems.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00165648).

REFERENCES

[1] H. W. Kwon, J.-W. Nam, J. Kim, and Y. K. Lee, "Generative adversarial attacks on fingerprint recognition systems," in *2021 International Conference on Information Networking (ICOIN)*, pp. 483–485, 2021.

[2] Y. Mittal, A. Varshney, P. Aggarwal, K. Matani, and V. K. Mittal, "Fingerprint biometric based access control and classroom attendance management system," in *2015 Annual IEEE India Conference (INDICON)*, pp. 1–6, 2015.
[3] M. Y. Lim, T. Y. Kim, J. E. Park, P. M. Hong, and Y. K. Lee, "A central point-based analysis for fingerprint liveness detection," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1307–1309, 2022.
[4] A. Toheed, M. H. Yousaf, A. Javed, *et al.*, "Physical adversarial attack scheme on object detectors using 3d adversarial object," in *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pp. 1–4, 2022.
[5] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, and S. K. Jha, "Directed adversarial attacks on fingerprints using attributions," in *2019 International Conference on Biometrics (ICB)*, pp. 1–8, 2019.
[6] K. Wang, "The advance of adversarial attack and defense," in *2022 International Conference on Applied Physics and Computing (ICAPC)*, pp. 262–265, 2022.
[7] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8571–8580, 2018.
[8] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 262–271, 2020.
[9] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, "Neighbor2neighbor: Self-supervised denoising from single noisy images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14781–14790, 2021.
[10] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," *arXiv:1803.04189*, 2018.
[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, 2014.
[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv:1706.06083*, 2017.
[13] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pp. 492–511, 2020.
[14] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1787, 2018.
[15] T. Dai, Y. Feng, D. Wu, B. Chen, J. Lu, Y. Jiang, and S.-T. Xia, "Dipdefend: Deep image prior driven defense against adversarial examples," in *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pp. 1404–1412, 2020.
[16] Y. Bakhti, S. A. Fezza, W. Hamidouche, and O. Déforges, "Ddsa: A defense against adversarial attacks using deep denoising sparse autoencoder," *IEEE Access*, vol. 7, pp. 160397–160407, 2019.
[17] V. Mura, L. Ghiani, G. L. Marcialis, F. Roli, D. A. Yambay, and S. A. Schuckers, "Livdet 2015 fingerprint liveness detection competition 2015," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6, 2015.
[18] H. Yoo, P. M. Hong, T. Kim, J. W. Yoon, and Y. K. Lee, "Defending against adversarial fingerprint attacks based on deep image prior," *IEEE Access*, vol. 11, pp. 78713–78725, 2023.
[19] P. Nahar, S. Tanwani, N. S. Chaudhari, *et al.*, "Fingerprint classification using deep neural network model resnet50," *International Journal of Research and Analytical Reviews (IJRAR)*, vol. 5, no. 4, pp. 1521–1537, 2018.
[20] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9446–9454, 2018.
[21] Y. Quan, M. Chen, T. Pang, and H. Ji, "Self2self with dropout: Learning self-supervised denoising from single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1890–1898, 2020.