Robustness of Face Recognition Adversarial Based on Image Denoising

Yi Gu

School of Computer Science, University of Birmingham, Birmingham, B15 2TT, United Kingdom

Abstract. Adversarial attacks have become an essential direction for security research in this area, as they generate adversarial samples to deceive face recognition systems. To defend against adversarial attacks, people need various methods to cope with the attacks, and the methods based on image denoising have widely attracted the attention of scholars for their High applicability and robustness. This paper aims to explore the application of three types of anti-attack defence in the field of face recognition by traditional filtering, JPEG compression denoising and deep learning denoising. The paper concluded the high efficiency of traditional filtering and JPEG compression denoising, with the lack of effect when defending against non-noise type attacking, however, deep learning noise reduction is far superior to traditional filtering or compression noise reduction in effect. On this basis, this paper further explores the insufficient attention to high-frequency noise denoising in the current study, and looks forward to the future research trend of combining image denoising and face recognition security, providing a reference for subsequent research.

1 Introduction

With the rapid development of deep learning techniques, the accuracy of face recognition systems has improved greatly, and it has been widely implied in the fields of Security Surveillance, Identity Authentication, Automated Driving, Medical Records Management, etc [1][2]. However, recent studies have found that face recognition systems have a serious vulnerability to adversarial attacks. The adversarial attack is a kind of deliberate perturbation that applies small, undetectable noise to input images [3], which leads to the face recognition system making the wrong classification. This kind of attack not only endangers the security and reliability of the recognition system, but may also bring about a series of real threats such as user privacy leakage, financial fraud, and illegal system intrusion. Thus, finding methods for improving the robustness of the face recognition system has become an important direction in the current research on facial recognition security.

For defence against attacks, researchers have proposed two kinds of modes, namely active and passive defence modes. The active defence mode emphasizes improving model performance, while the passive defence mode focuses on reducing perturbations [4]. Active

yxg479@student.bham.ac.uk

defence includes adversarial training, input transformation, feature compression, adversarial example detection, and image denoising, among others [5]. Among them, image denoising, as a defence method operating at the image input level, has attracted increasing attention from researchers due to its model-agnostic and attack-agnostic characteristics. The core concept is to apply denoising techniques as a preprocessing step to input images [3], aiming to eliminate or mitigate adversarial noise interference, thereby recovering the authentic features of the image and enhancing the stability and accuracy of recognition systems.

This article will start with the anti-attack defence of the face recognition system. The system sorts out the principles and applications of the current major noise-denoising technology, analyzes the calculation overhead and efficiency of various methods, and summarizes the defence effects and applicable scenarios of different methods on the anti-attack samples. On this basis, the shortcomings of the current research on high-frequency noise denoising research [6] will be further explored, and the research trends in related fields will be looked forward to in the future.

2 An Overview of Techniques for Image Denoising

Currently, image denoising technology can be roughly divided into three categories, as shown in Table 1 and Table 2.

2.1 Robust Face Recognition with Sparse Representation

The first category is the robust face recognition method proposed by Wright et al., which uses sparse representation to achieve efficient face recognition in a noisy environment [7]. This method includes Gaussian filtering, median filtering, etc. These techniques use the local statistical characteristics of the image to smooth out noise. Traditional filtering methods are simple to implement but are limited in their effectiveness against complex noise. As a result, they are often used as preprocessing methods.

2.2 JPEG Noise Reduction Method

The second type is the JPEG noise reduction method, where the image is compressed into JPEG format, which can also help reduce noise to some extent. By performing JPEG compression on the image and removing some high-frequency information [8], noise interference can be minimized. This method can sometimes provide a certain degree of defense against adversarial attacks [9]. However, JPEG noise reduction mainly works by losing some information, and it cannot effectively handle complex adversarial attacks. Therefore, while JPEG compression is occasionally used as a fast preprocessing method, it does not fundamentally solve the problem of adversarial attacks.

2.3 Deep Learning-based Denoising Methods

In the third category, the deep learning-based denoising method proposed by Zhang et al. uses neural networks such as CNN, RNN, and other models to learn noise distribution and efficiently denoise the image [10]. Compared with the methods mentioned above, deep learning techniques offer stronger adaptability and expression capabilities but come with higher computational costs. While deep learning-based methods improve the robustness of facial recognition systems against attacks to some extent, they still face challenges such as large computational overhead, insufficient generalization to unknown attacks, and reduced

recognition accuracy due to information loss. These issues require further research and improvement.

Table 1 Categories of Optimization Methods

Category	Optimization Direction	Main Methods
Data-level Optimization	Data Preprocessing [11]	Gaussian filtering, median filtering, JPEG compression, deep learning denoising
	Data Augmentation [12]	Random cropping, rotation, color jitter
Model-level Optimization	Regularization Methods [13]	L1/L2 regularization, Dropout, adversarial training

Table 2 Characteristics of Commonly Used Denoising Methods

Method Category	Representative Algorithms	Main Advantages	Main Disadvantages	Applicable Scenarios
Traditional Filtering	Gaussian Filter / Median Filter	Simple and efficient	Poor performance against complex attacks	Low-risk scenarios
JPEG Compression Denoising	JPEG Compression	Fast and easy; reduces high-frequency noise	Over-compression may degrade image quality; limited effectiveness	The preprocessing stage against adversarial attacks
Deep Learning Denoising	DnCNN / U-Net, etc.	Strong denoising performance	High training cost; poor generalization to unknown attacks	High-security requirement scenarios

3 Case Study

This study focuses on defending against adversarial attacks targeting image recognition systems, specifically investigating denoising-based defenses against digital adversarial attacks. The core objective is to explore how image denoising techniques can effectively eliminate the interference of attacks on facial recognition systems, while comprehensively evaluating the defence performance and adversarial robustness of different denoising methods.

3.1 Deep Learning-Based Denoising Approaches

According to previous research, deep learning-based denoising approaches (such as DAE and DUNET) can reduce adversarial noise [5]; however, when integrated with the original network, they may become more prone to perturbations. Liao et al. proposed a novel deep learning-based autoencoder that effectively addresses this issue [14]. Experimental results are presented in Table 3(adapted from [14]).

Table 3 Experimental Results of Deep Learning-Based Denoising Methods (Adapted from [14])

Defense	Clean	WhiteTestSet (ϱ = 4)	WhiteTestSet (ϱ = 16)	BlackTestSet (ϱ = 4)	BlackTestSet (ϱ = 16)	Denoising Loss (ϱ = 4)	Denoising Loss (ϱ = 16)
NA	76.70%	14.50%	14.40%	61.20%	41.00%	0	0.0177
DAE	58.30%	51.40%	36.70%	55.90%	48.80%	0.036	0.0359
DUNET	75.30%	20.00%	13.80%	67.50%	55.70%	0.015	0.014
PGD	75.30%	20.00%	13.80%	67.50%	55.70%	-	-
ensV3	76.90%	69.80%	58.00%	72.40%	62.00%	-	-
FGD	76.10%	73.70%	67.40%	74.30%	71.80%	-	-
LGD	76.20%	75.20%	69.20%	75.10%	72.20%	-	-
CGD	74.90%	75.80%	73.20%	74.50%	71.10%	-	-

3.2 Traditional Filtering Methods

According to Dziugaite's research [9], anti-attack noise is an obvious "noise", so the effect is better when using a convolutional denoising strategy, that is, traditional filtering. Among them, by comparing the four strategies of non-local mean, bilateral filtering, mean filtering and median filtering, it is concluded that the non-local mean has the best effect. However, compared with JPGE denoising, traditional filtering methods do not have special treatment for high-frequency noise that frequently occurs in noise.

3.3 JPEG Compression Denoising

In the compression denoising method represented by JPGE compression denoising, researchers have shown that using JPGE compression can remove disturbed noise, but cannot effectively remove disturbed noise [15] [16], and it is possible to further reduce the recognition accuracy. For specific data, refer to Table 4(adapted from [15]).

Table 4 Performance of JPEG Compression Denoising (Adapted from [15])

Attack Method	Model	JPEG Quality (q)	Accuracy (%)	Success Rate (%)	Notes
None	N/A	∞ (no JPEG)	70.7	N/A	Baseline model without JPEG defense
FGSM (ε =1/255)	Untargeted	∞ (no JPEG)	12.2	N/A	Fast Gradient Sign Method (FGSM) attack
FGSM (ε =1/255)	Untargeted	75	36.8	N/A	JPEG quality defense impact on FGSM
FGSM (ε =1/255)	Untargeted	50	45.8	N/A	JPEG compression reduces FGSM attack
FGSM (ε =3/255)	Untargeted	∞ (no JPEG)	4.8	N/A	No JPEG defense
FGSM (ε =3/255)	Untargeted	75	14.6	N/A	JPEG compression reduces attack success

FGSM (ε =3/255)	Untargeted	50	23.1	N/A	JPEG compression reduces attack success
FGSM (ε =3/255)	Untargeted	25	35.4	N/A	JPEG compression reduces attack success
I-FGSM (ε =1/255)	Untargeted	∞ (no JPEG)	1.6	N/A	Iterative FGSM attack
I-FGSM (ε =1/255)	Untargeted	75	34.6	N/A	JPEG compression slightly reduces attack
I-FGSM (ε =1/255)	Untargeted	50	47.4	N/A	JPEG compression reduces attack success
I-FGSM (ε =1/255)	Untargeted	25	52	N/A	JPEG compression reduces attack success
I-FGSM (ε =3/255)	Untargeted	∞ (no JPEG)	0.2	N/A	No JPEG defense
I-FGSM (ε =3/255)	Untargeted	75	9.5	N/A	JPEG compression reduces attack success
I-FGSM	Untargeted	50	30.4	N/A	JPEG compression

(ε =3/255)					reduces attack success
I-FGSM (ε =3/255)	Untargeted	25	44.2	N/A	JPEG compression reduces attack success
Targeted I-FGSM	Targeted	∞ (no JPEG)	99.2	0.1	Targeted attack success rate with no JPEG
Targeted I-FGSM	Targeted	75	0.96	0.02	JPEG compression drastically reduces attack
Targeted I-FGSM	Targeted	50	0.01	0.07	JPEG compression reduces attack success
Targeted I-FGSM	Targeted	25	0.001	0.69	JPEG compression reduces attack success

3.4 Limitations of Denoising Against Non-Noise Attacks

According to research, there are a large number of non-noise variants [17] against the face [17], such as occlusion and distortion of the face. As shown in Figure 1, attacks that do not belong to noise, so it is difficult to defend with denoising methods. Attackers can use a variety of attack methods to attack jointly.

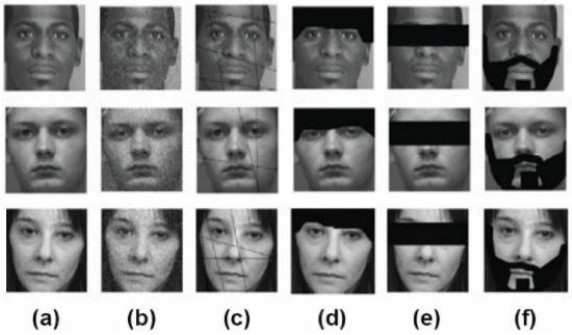

Figure 1 Other Types of Facial Adversarial Attacks [17]

4 Discussion

This study provides an in-depth analysis of the defence against adversarial attacks on image recognition systems, especially face recognition systems and explores the defence effect based on image denoising techniques. It is found that deep learning-based denoising schemes (e.g., DAE and DUNET) can effectively reduce the interference of adversarial attacks on face recognition systems to a certain extent, especially in the face of white-box and black-box attacks with more significant performance. The experimental results show that the denoising loss of the deep learning denoising method is low, and the defence effect is still more stable under different attack strengths.

However, traditional filtering strategies such as non-local mean filtering also show some advantages in dealing with noise against attacks, especially in removing obvious noise perturbations. Nevertheless, traditional filtering methods are not specifically optimized for high-frequency noise, thus limiting their effectiveness in the face of complex attacks.

In addition, the study also reveals that other types of adversarial attacks (e.g., masking or distorting faces) besides noise-based attacks also pose a threat to the system, which cannot be effectively defended by conventional de-noising methods, and require that future defence schemes are not only limited to noise suppression but also need to incorporate more complex feature processing and a joint defence strategy for multiple attacking methods.

Although this study explores the defence methods based on image-denoising techniques to cope with adversarial attacks by summarizing the results of previous research, the proposed conclusions mainly rely on the data and analysis in the existing literature due to the lack of practical experimental verification. Therefore, future research should focus on the reproduction and validation of actual experiments to further evaluate the defence effects of different denoising methods in real scenarios.

5 Conclusion

In this paper, we mainly focus on the defence methods of image recognition systems, especially face recognition systems, in the face of adversarial attacks, and investigate the defence effect based on image denoising techniques. By analyzing previous research results, this paper summarizes several common denoising methods, including deep learning-based denoising methods (e.g., DAE and DUNET), traditional filtering strategies (e.g., nonlocal mean filtering), and compression denoising methods (e.g., JPEG compression). The study explores the performance of different methods in dealing with adversarial attacks of different strengths by comparing their defense effectiveness.

The main findings of this paper include: first, deep learning-based denoising methods (e.g., DAE and DUNET) are able to effectively inhibit the impact of adversarial attacks on face recognition systems to a certain extent, especially when dealing with white-box and black-box attacks, and show a more significant defence effect. Secondly, traditional non-local mean filtering performs well in removing noise perturbations, although it fails to optimize specifically for high-frequency noise. In contrast, JPEG compression, as a means of denoising, reduces the success rate of attacks but may lead to a decrease in recognition accuracy when dealing with high-intensity attacks. Furthermore, it was found that in addition to noise-based attacks, there are also non-noise-based attacks (e.g., occlusion and distortion) that pose additional challenges for defense, which cannot be defended by conventional denoising methods.

In terms of outlook, future research can build on this study to conduct experimental replications to evaluate the effectiveness of different denoising methods in real-world environments. Meanwhile, combining denoising techniques with other defence mechanisms, such as multimodal defence strategies, feature selection, or Generative Adversarial Networks (GANs), may further improve the robustness of image recognition systems. In addition, research on defense against non-noise-based attacks is also an important future direction, and it is worth exploring in depth how to combine denoising methods with other techniques to provide image recognition systems with stronger defense capabilities.

In summary, this study not only provides theoretical support for the defence of image recognition systems against attacks but also provides a valuable direction for further research and practice in the future.

References

1. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: 'A survey of deep learning techniques for autonomous driving', J. Field Robot., 2020, 37, (3), pp. 362–386
2. Zhao, Z. Q., Huang, D. S., Sun, B. Y.: 'Human face recognition based on multi-features using neural networks committee', Pattern Recognit. Lett., 2004, 25, (12), pp. 1351–1358
3. Goodfellow, I. J., Shlens, J., Szegedy, C.: 'Explaining and Harnessing Adversarial Examples', 2014. Available at: http://arxiv.org/abs/1412.6572
4. Wang, J., Wang, C., Lin, Q., Luo, C., Wu, C., Li, J.: 'Adversarial attacks and defenses in deep learning for image recognition: A survey', Neurocomputing, 2022, 514, pp. 162–181
5. Gu, S., Rigazio, L.: 'Towards Deep Neural Network Architectures Robust to Adversarial Examples', 2014. Available at: http://arxiv.org/abs/1412.5068
6. Zhang, Z., Zeng, Y., Liu, Q., Zhou, S.: 'Towards a Novel Perspective on Adversarial Examples Driven by Frequency', 2024. Available at: http://arxiv.org/abs/2404.10202
7. Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., Ma, Y.: 'Robust face recognition via sparse representation', IEEE Trans. Pattern Anal. Mach. Intell., 2009, 31, (2), pp. 210–227
8. Wallace, G. K.: 'The JPEG still picture compression standard', 1992
9. Dziugaite, G. K., Ghahramani, Z., Roy, D. M.: 'A study of the effect of JPG compression on adversarial images', 2016. Available at: http://arxiv.org/abs/1608.00853
10. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: 'Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising', IEEE Trans. Image Process., 2017, 26, (7), pp. 3142–3155

11. Rakin, A. S., He, Z., Gong, B., Fan, D.: 'Blind Pre-Processing: A Robust Defense Method Against Adversarial Examples', 2018. Available at: http://arxiv.org/abs/1802.01549
12. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: 'Mitigating Adversarial Effects Through Randomization', 2017. Available at: http://arxiv.org/abs/1711.01991
13. Szegedy, C., et al.: 'Intriguing properties of neural networks', 2013. Available at: http://arxiv.org/abs/1312.6199
14. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: 'Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser'. Available at: https://github.com/lfz/Guided-Denoise
15. Shin, R., Song, D.: 'JPEG-resistant Adversarial Images'. Available at: https://github.com/tensorflow/models/tree/master/research/slim
16. Das, N., et al.: 'Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression', 2017. Available at: http://arxiv.org/abs/1705.02900
17. Vakhshiteh, F., Nickabadi, A., Ramachandra, R.: 'Adversarial Attacks against Face Recognition: A Comprehensive Study', IEEE Access, 2021, 9, pp. 92735–92756