Artificial Immune System of Secure Face Recognition Against Adversarial Attacks

Min Ren[1†], Yunlong Wang[2†], Yuhao Zhu[3], Yongzhen Huang[1*], Zhenan Sun[2], Qi Li[2], Tieniu Tan[2]

[1]School of Artificial Intelligence, Beijing Normal University, Beijing, China.
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[3]Postgraduate Department, China Academy of Railway Sciences, Beijing, China.

*Corresponding author(s). E-mail(s): huangyongzhen@bnu.edu.cn;
Contributing authors: renmin@bnu.edu.cn;
yunlong.wang@cripac.ia.ac.cn; zhuyuhao@rails.cn; znsun@nlpr.ia.ac.cn;
qli@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn;
[†]These authors contributed equally to this work.

Abstract

Deep learning-based face recognition models are vulnerable to adversarial attacks. In contrast to general noises, the presence of imperceptible adversarial noises can lead to catastrophic errors in deep face recognition models. The primary difference between adversarial noise and general noise lies in its specificity. Adversarial attack methods give rise to noises tailored to the characteristics of the individual image and recognition model at hand. Diverse samples and recognition models can engender specific adversarial noise patterns, which pose significant challenges for adversarial defense. Addressing this challenge in the realm of face recognition presents a more formidable endeavor due to the inherent nature of face recognition as an open set task. In order to tackle this challenge, it is imperative to employ customized processing for each individual input sample. Drawing inspiration from the biological immune system, which can identify and respond to various threats, this paper aims to create an artificial immune system (AIS) to provide adversarial defense for face recognition. The proposed defense model

incorporates the principles of antibody cloning, mutation, selection, and memory mechanisms to generate a distinct "antibody" for each input sample, wherein the term "antibody" refers to a specialized noise removal manner. Furthermore, we introduce a self-supervised adversarial training mechanism that serves as a simulated rehearsal of immune system invasions. Extensive experimental results demonstrate the efficacy of the proposed method, surpassing state-of-the-art adversarial defense methods. The source code is available here, or you can visit this website: *https://github.com/RenMin1991/SIDE*

Keywords: Adversarial Defense, Face Recognition, Artificial Immune System, Self-supervised Adversarial Learning

1 Introduction

Deep learning-based feature extractors have garnered significant achievements across diverse domains, notably in image classification [1–7], object detection [8, 9], and semantic segmentation [10, 11]. This can be attributed to their capability of executing non-linear mappings from raw data to high-dimension features. Despite the powerful expressive capabilities of deep learning models, their susceptibility to adversarial attacks undermines their reliability and degrades their security [12–16]. Hence, numerous researchers have directed their attention towards adversarial defense techniques for deep learning models [17–24].

As a prevalent and widely adopted application of deep learning technology, deep learning based face recognition [25–33] have demonstrated their ability to surpass human performance in both verification and identification scenarios. The domains where face recognition is applied, such as finance and border control, typically impose stringent requirements on security. Nevertheless, the security of deep learning-based face recognition systems is greatly degraded by the inherent fragility of deep learning models against adversarial attacks. Whether these attacks stem from the digital domain [34] or are directly imposed in the physical domain [35], they effortlessly exploit vulnerabilities within face recognition models, leading to catastrophic errors. Thus, the pursuit of adversarial defense methods in the domain of face recognition not only holds theoretical significance but also represents a pressing technological imperative driven by practical application demands.

In fact, deep learning models demonstrate remarkable robustness against common types of noise, such as Gaussian and salt-and-pepper noise. Their susceptibility to adversarial noises can be attributed to the specificity possessed by these noises. Adversarial attack methods generate noises that are specifically tailored to the characteristics of the target image and recognition model. As a consequence, distinct samples and recognition models can lead to the creation of specific patterns of adversarial noise, thus imposing considerable challenges when devising effective mechanisms for adversarial defense.

Facial recognition can be applied to both close-set scenarios, where all identities are available, and open-set scenarios, where the identities encountered during testing are

unavailable. This challenge poses great difficulties for the face recognition task due to the fact that face recognition tasks involve dealing with a substantially larger number of unique identities, especially in open-set scenarios. Hence, adversarial samples in the context of face recognition demonstrate a great degree of diversity, complexity, and unpredictability in their adversarial noise patterns. The specificity inherent in adversarial noises amplifies their detrimental impact, thereby intensifying the vulnerability of face recognition models to adversarial attacks.

In response to this challenge, we propose a novel adversarial defense model for face recognition. The proposed model provides individualized manners for removing adversarial noises, thus endowing each facial image with a tailored and specific defense against adversarial noises. This approach draws inspiration from the biological immune system, which exhibits powerful attributes such as self-learning, dynamic adaptation, and memory capabilities [36]. The immune system can generate specific antibodies against various viruses through processes of antibody cloning, mutation, selection, and memory. These mechanisms enable the production of effective antibodies that bind specifically to antigens, thus achieving effective immunity.

In the proposed method, adversarial noises can be analogous to antigens, while the noise removal ways are analogous to antibodies. When devising the structure of antibodies, the perturbation inactivation methodology of PIN [37] is assimilated. This method is an adversarial defense approach that utilizes eigenvectors of facial images to filter noise and restoring essential facial features. However, PIN fails to consider the specificity of adversarial noises during noise removal, making it challenging to effectively differentiate between the harmful information introduced by adversarial noises and the inherent information of the face itself. As a result, striking a balance between adversarial noise removal and facial sample restoration becomes difficult. This issue is also encountered by most adversarial defense methods based on noise removal. In contrast, we propose an artificial immune system that provides customized noise removal ways for facial samples. This is achieved through the generation, cloning, mutation, and memory mechanisms of antibodies within the immune system. The proposed defense model consists of three essential components: the antigen analyzer, the antibody generator, and the memory module. The antigen analyzer is responsible for analyzing the characteristics of adversarial noises, whereas the antibody generator emulates the processes of antibody cloning, mutation, and selection to optimize the antibodies. Concurrently, the memory module is employed to store patterns of adversarial noises during the optimization process.

In addition, we propose a self-supervised adversarial training mechanism that collaborates with the aforementioned adversarial defense model. This mechanism integrates a momentum-updated siamese network of the adversarial defense model to generate on-the-fly adversarial samples. Self-supervised adversarial training can offer more precise guidance for the process of antibody selection, leading to a gradual enhancement in their defense capabilities. This process is analogous to how the immune system continuously enhances its immune capabilities through repeated confrontations with viral invasions.

The main contributions of this paper can be summarised as follows:

- In this paper, we introduce a novel face recognition adversarial defense model based on the principles of specific immunity. By emulating the intricate process of specific immune evolution observed in biological systems, this model provides tailored denoising ways for individual input facial images. This model effectively mitigates the challenges posed by the specificity of adversarial noise to face recognition models, enabling robust and reliable recognition performance.
- We introduce a self-supervised adversarial training mechanism that contributes to the selection of "antibodies" within the proposed adversarial defense model. This mechanism facilitates iterative refinement through self-adversarial training, empowering the model to enhance its defensive capabilities against adversarial noises.
- The effectiveness of the proposed adversarial defense method has been experimentally validated across diverse types of adversarial attacks and multiple datasets. The experimental results demonstrate its superior performance compared to existing state-of-the-art adversarial defense methods. Furthermore, we offer a comprehensive experimental analysis, providing valuable insights into the effectiveness and robustness of the proposed method.

The remainder of this paper is organized as follows: Section 2 presents a brief literature review of the related work. The proposed adversarial defense model and the self-supervised adversarial training mechanism are described in detail in Section 3. The configurations and results of experiments are presented in Section 4. Finally, the conclusion of this paper is summarized in Section 5.

2 Related Work

2.1 Deep Learning Based Face Recognition

Deep learning-based face recognition has achieved remarkable advancements in recent years. The pioneering work of CNN-based face representation was introduced by Taigman et al. [25] and Yi et al. [26]. In these studies, face recognition is regarded as a multi-class classification challenge. To address this, deep convolutional neural network (CNN) models are initially implemented to acquire features from extensive datasets containing multiple identities. However, due to the fact that face recognition is an open-set task, meaning testing identities that are different from the ones used for training, there is a significant difference between face recognition and image classification tasks. As a result, researchers have increasingly tended to model it as a metric learning task, aiming to learn a feature space mapping model with strong discriminative power through constraints imposed on the feature space. In the pursuit of obtaining a 128-D face embedding representation, Schroff et al. utilized a triplet loss function in their research [27]. To further enhance feature embedding, numerous methods have been proposed, such as SphereFace [28], CosFace [29], ArcFace [30], among others. In addition, to facilitate the deployment of facial recognition models, researchers have also focused on the study of lightweight face recognition models [38, 39].

In spite of the considerable progress made in this field, deep learning models utilized for face recognition remain susceptible to adversarial attacks, as indicated by previous

studies [34, 35, 40]. This vulnerability presents a profound concern and jeopardizes the overall security of face recognition systems.

2.2 Adversarial Attack

The adversarial attack technique for computer vision tasks has become a prominent area of research. Szegedy et al. [12] is the first to demonstrate the vulnerability of deep neural networks to adversarial noises. Since then, numerous methods for adversarial attacks have been proposed. Goodfellow et al. [13] introduced an efficient single-step attack method called FGSM, which is based on gradient calculations. DeepFool [41] aims to identify the nearest decision boundary in order to confuse the model. C&W [42] addresses the joint optimization of the objective function and the scale of noises. Projected gradient descent (PGD) [43] iteratively applies the gradient signal of deep learning models, which is the most powerful first-order adversarial attack method [43]. Su et al. [44] propose an intriguing approach that confuses deep learning models by altering just a single pixel in the image. Additionally, there have been reports on the generalization of adversarial noises [45–49]. Universal adversarial noises based attack methods are proposed in several studies [14, 50–52].

Recently, there have been reports of targeted adversarial attack techniques specifically tailored for face recognition systems. Dong et al. [34] introduce a decision-based adversarial attack method for face recognition. The rapidly evolving field of transferable facial adversarial attack techniques has provided additional avenues for black-box facial adversarial attacks [53, 54]. Furthermore, instances of physical domain adversarial attacks have also been documented in the literature. Sharif et al. [55] presented a systematic approach for generating physically feasible attacks by printing a pair of eyeglass frames. Another method, known as sticker attack, was proposed by Komkov et al. [35], which involves using a specially designed rectangular paper sticker to deceive the face recognition system. Recently, there have been further developments in sticker-based adversarial attacks on faces. Yang et al. [56] introduce a sticker generation method that can adhere to the three-dimensional shape of faces. These real-world attacks pose new challenges to the adversarial defense for face recognition systems.

2.3 Adversarial Defense

The extant adversarial defense methods can be broadly categorized into two distinct groups. The first group encompasses methods that strive to enhance the robustness of neural networks against adversarial examples. These methods concentrate on fortifying the network's capacity to withstand noises and maintain accurate predictions in the presence of such malicious input. On the other hand, the second group involves methodologies that aim to eliminate the adversarial noises from the adversarial samples prior to their presentation to the recognition model. This category of approaches focuses on cleansing the input data by removing the embedded harmful alterations, thereby reducing the potential impact of adversarial attacks.

A prevalent strategy of the first type involves training neural networks using adversarial examples [13, 17–19, 57]. This strategy is straightforward and aims to

enhance the network's resistance against adversarial attacks. To improve the robustness against gradient-based attacks, several learning strategies have been proposed. Ross et al. [20] trained models with input gradient regularization. Other techniques, such as network distillation [58], region-based classifier [59], generative model [60, 61], and self-supervised learning [62] have also been adopted to enhance model robustness. Rakin et al. [63] introduced a trainable randomness method for adversarial training to improve robustness. A novel loss function for adversarial defense is proposed for adversarial defense in [64]. Mustafa et al. [65] achieved elevated robustness by constraining the hidden space of deep neural networks. Zhong et al. [66] utilized margin-based triplet embedding regularization to train the recognition model. Cazenavette et al. [67] aimed to enhance the adversarial robustness of CNNs by reframing each layer as a sparse coding model. Jin et al. [68] present an approach for analyzing the noise pattern by Taylor expansion. However, these methods often exhibit poor generalization to adversarial noise that does not appear in the training set, as verified by our experiments in this paper. This is because the patterns of adversarial noise are more complex and diverse in face recognition tasks, and relying solely on adversarial training is insufficient to cope with them.

The other type of approaches are devised to eliminate the adversarial noises prior to the recognition model's processing [37, 69–73]. Das et al. [69] proposed the application of JPEG compression to remove these noises. In a study by Guo et al. [70], image quilting and total variation minimization (TVM) were assessed as possible techniques for this purpose. Meng et al. [71] introduced a two-pronged defense strategy to effectively eliminate adversarial noises. Liao et al. [72] incorporated the U-Net [74] as a denoising module, thereby enabling the removal of adversarial noises. The work presented in [73] employed PixelCNN [73] to transform adversarial examples into clean images. Bai et al. [75] further improved the defense performance by incorporating Hilbert scan into PixelCNN. Dezfooli et al. [76] and Sun et al. [77] utilized sparse coding to reconstruct image patches. Gupta et al. [78] attempted to identify the most influential regions of an image for reconstruction. Xie et al. [79] employed a self-attention layer to recover the original information within the feature space. Zhou et al. [80] utilized self-supervised learning to eliminate adversarial noise in the class activation feature space. PIN [37] utilizes eigenvectors of facial images to filter noise and restoring essential facial features. However, it is worth noting that most of these methods were primarily developed for general image classification tasks. Furthermore, they are unable to provide specific noise removal ways, rendering them unsuitable for face recognition tasks.

2.4 Algorithms Inspired by the Immune System

The biological immune system is an evolved defense mechanism in vertebrates to protect the organism from the invasion of "non-self" entities such as pathogens. Due to its superior characteristics, including self-learning, memory mechanisms, and dynamic adaptability [36], the biological immune system has provided abundant biomimetic inspiration for solving various problems. Artificial immune systems (AIS) [81] construct algorithms by simulating the functions, principles, and models of the biological immune system.

Among numerous immunological theories, the clonal selection theory [36] has provided significant inspiration for the development of computer algorithms [81–85]. The clonal selection theory explains the fundamental characteristics of adaptive immune response under antigen stimulation. Its basic concept is that only those cells capable of recognizing antigens are selected for proliferation, while those incapable of antigen recognition are not selected. The selected cells undergo proliferation and mutation processes to enhance their affinity.

Drawing inspiration from the principles of the clonal selection theory, researchers have put forth numerous algorithms that have yielded fruitful research outcomes in domains such as dynamic programming [84] and multi-objective optimization [85]. Clonal selection algorithms commonly incorporate the following essential components: affinity calculation, which quantifies the quality of antibodies; selection, which involves the screening of existing antibodies; cloning, which encompasses the replication of antibodies; mutation, which entails modifying the structure of antibodies; and memory, which involves the storage of information regarding antibodies [81].

This paper stands as a pioneering effort that amalgamates the clonal selection algorithm with deep learning methodologies for adversarial defense. We propose a novel approach to address the specificity and complexity of adversarial noise in face recognition tasks by implementing the aforementioned components of the clonal selection algorithm. This approach has successfully achieved adversarial defense tailored for face recognition.

3 Methodology

This section provides a comprehensive description of the proposed defense method. Firstly, we present the preliminaries of the method, encompassing symbol representation and the definition of antibodies. Building upon this foundation, we introduce the structure and overview of the proposed model. Subsequently, we delve into the defense model optimization and the self-supervised adversarial defense training. Finally, we provide the implementation details of the method.

3.1 Preliminaries

A face recognition model can be regarded as a mapping function that transforms facial images into the feature space. Given a facial image represented as $x \in \mathbb{R}^{C \times H \times W}$, the deep feature f can be extracted using a face recognition model denoted as F:

$$f = F(x) \tag{1}$$

Adversarial samples are generated by the adversarial attack method according to the facial image x and the face recognition model F:

$$x_{adv} = Attacker(x, F) \tag{2}$$

where $Attacker$ is the adversarial attack method, x_{adv} is the adversarial sample. The adversarial noise is the difference between x and x_{adv}. In this paper, we propose an

adversarial defense model, denoted as D, in this paper for recovering x from x_{adv}:

$$x_{recon} = D_\theta(x_{adv}) \tag{3}$$

where θ is the parameters of the proposed model.

Definition of Antibody:

In the context of adversarial defense, the adversarial noises in adversarial samples can be viewed as antigens, while noise removal methods can be seen as antibodies. The fundamental motivation behind the design of antibody form lies in facilitating the analysis and removal of noise. Given the intricate and diverse nature of adversarial noises in facial recognition, modeling the distribution of clean facial images in pixel space serves as a foundation for mitigating the challenges associated with noise analysis. By acquiring a comprehensive understanding of the underlying distribution of clean facial images, we can establish a solid basis for reducing the complexity involved in noise analysis and consequently lowering the difficulty in effectively removing adversarial noises.

In the realm of facial image distribution modeling, researchers have already made significant strides, providing valuable insights for further investigation. Among these noteworthy contributions, EigenFace [86] stands out as a pioneering and groundbreaking approach. Following the lead of EigenFace, we can model the distribution of facial images in pixel space by the eigenvectors that characterize the data distribution. Each eigenvector represents a distinct dimension in pixel space, and the components of facial images across different eigenvectors express various types of features. This provides a favorable foundation for us to design noise removal methods, namely antibodies. Therefore, we define antibodies using eigenvectors: an antibody is a composition of eigenvectors:

$$a = \{e_1, e_2, ..., e_n\} \tag{4}$$

where a is an antibody, $e_i \in \mathbb{R}^d$, $d = C \times H \times W$ is an eigenvector of the facial images, n is the number of eigenvectors of the antibody, and $0 < n < d$.

The eigenvectors comprising antibodies have the ability to selectively filter features of a facial image, retaining the characteristics corresponding to these eigenvectors:

$$\alpha = E^T(x_{adv}^{flat} - x_{mean}) \tag{5}$$

where $E \in \mathbb{R}^{d \times n}$ is a matrix composed of the eigenvectors from an antibody, wherein each column represents an eigenvector, $x_{adv}^{flat} \in \mathbb{R}^d$ is obtained by reshaping x_{adv} into a flattened vector, $x_{mean} \in \mathbb{R}^d$ is the mean vector of the facial samples in the pixel space. $\alpha \in \mathbb{R}^n$ represents the component of the input image along the eigenvectors of the antibody, encompassing only the characteristic information associated with these eigenvectors. After obtaining α, the facial image can be reconstructed:

$$x_{recon} = E\alpha^T + x_{mean} \tag{6}$$

By performing the aforementioned processing steps, the reconstructed image will only contain the features corresponding to the eigenvectors of the antibody, while the
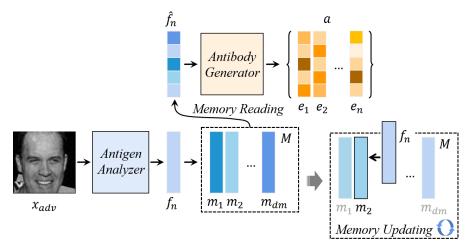
8

Fig. 1 The architecture of the proposed adversarial defense method. The proposed adversarial defense model encompasses three key components: the antigen analyzer, the antibody generator, and the memory module.
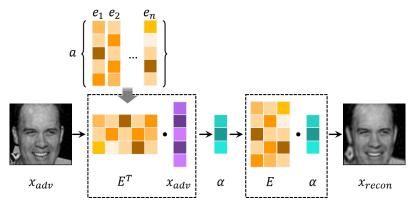

Fig. 2 The process of using antibodies for noise removal (omitting x_{mean} for brevity). The eigenvectors comprising antibodies have the ability to selectively filter facial features in an image, retaining the characteristics corresponding to these eigenvectors while removing the remaining information.

features corresponding to eigenvectors not present in the antibody are removed. For different adversarial samples, employing distinct antibodies enables the removal of different features, thereby facilitating a tailored and specific noise removal manner for each input sample.

3.2 Model Architecture

The proposed adversarial defense model encompasses three key components: the antigen analyzer, the antibody generator, and the memory module, as shown in Fig. 1.

In order to provide specific treatment for each input adversarial face image, it is essential to conduct an effective analysis of these images. To this end, we adopt a deep neural network, serving as a noise analyzer for adversarial samples, namely the antigen

analyzer. Given an adversarial sample x_{adv}, the antigen analyzer, denoted as H, takes the adversarial sample as its input and produces the noise feature $f_n \in \mathbb{R}^{d_n}$ as output:

$$f_n = H(x_{adv}) \tag{7}$$

The noise information of the input image is contained in f_n.

Inspired by the memory mechanisms in the immune system, a memory module is incorporated into the proposed defense model. This memory module is used to store the noise patterns of adversarial samples, enabling the defense model to explicitly model vulnerabilities in the face recognition model and guide the generation of "antibodies". The memory module, represented by the noise feature matrix $M \in \mathbb{R}^{d_n \times d_m}$, serves as a repository of noise patterns, with d_m denoting the number of memory items. Before generating antibodies, f_n is utilized to retrieve noise features from the memory module M. Subsequently, a self-attention mechanism is employed to aggregate the retrieved noise features as the input of antibody generator $\hat{f}_n$. The detailed process of feature retrieval and aggregation, as well as the updating of the memory module, will be described in Section 3.3.

The antibody generator initially maps the aggregated noise feature $\hat{f}_n$ onto the eigenvector selection space through a feature mapping layer:

$$f_e = Sigmoid(G(\hat{f}_n)) \tag{8}$$

where G is the mapping layer, Sigmoid function is employed to normalize the output within the range of $(0, 1)$. The dimension of f_e is the same as the number of eigenvectors, with each dimension corresponding to the probability of including a specific eigenvector in the antibody. Subsequently, an antibody, which is a combination of eigenvectors, is obtained by sampling based on f_e. After obtaining the antibody, a tailored and specific noise removal manner for each input sample can be implemented as described in Section 3.1.

3.3 Memory Mechanism

In the immune system, the memory mechanism enable the retention of antigenic information, empowering it to swiftly generate effective antibodies upon encountering similar antigens. Taking inspiration from this phenomenon, we leverage a memory module to store the noisy features captured during the training process.

Memory Reading:
Upon obtaining the noise feature f_n, we employ it as a query to retrieve the items in the memory module $M \in \mathbb{R}^{d_n \times d_m}$:

$$r_i = \frac{m_i f_n^T}{||m_i|| \; ||f_n||}, \; i = 1, 2, ..., d_m \tag{9}$$

where $m_i \in \mathbb{R}^{d_n}$ is the i-th item of M, d_m is the number of memory items, r_i is the similarity between the query f_n and the i-th item. Subsequently, by aggregating the

10

memory items by soft-attention, we can accomplish memory retrieval and obtain the output of the memory module:

$$\hat{f}_n = \sum_{i=1}^{d_m} r_i m_i \tag{10}$$

Memory Updating:

During the training process, continuous updating of the memory model is necessary to adapt to changes in model parameters. To achieve this, we first identify the memory item that is most similar to f_n:

$$i^* = \arg\max_i \frac{m_i f_n^T}{||m_i|| \, ||f_n||}, \ i = 1, 2, ..., d_m \tag{11}$$

Afterwards, the memory module is updated via a moving average of f_n:

$$m_i \leftarrow \begin{cases} m_i & i \neq i^*; \\ \epsilon m_i + (1 - \epsilon) f_n & i = i^*. \end{cases} \tag{12}$$

where $\epsilon \in (0, 1)$ is the decay rate of memory updating. This means that the model updates only the memory item that is most similar to f_n each time while leaving the other items unchanged. As new adversarial noises continue to emerge, the memory module can store previous noise features.

3.4 Model Optimization

To train the defense model presented in the previous subsection, this subsection introduces the method for optimizing the model. Drawing inspiration from the immune system's generation and selection process of antibodies, we propose an antibody optimization approach that includes the following elements: antibody affinity, cloning and mutation, antibody screening, and model updating. In this subsection, we introduce each of these elements in detail and then present the antibody optimization algorithm based on them.

Antibody affinity:

In immunology, antibody affinity refers to the strength of the bond between an antibody and an antigen. The higher the affinity, the better the antibody's ability to neutralize the antigen and the more effective the immune response. We borrow this concept to measure the effectiveness of noise removal in adversarial defense. The definition of antibody affinity is as follows:

$$s(a) = \lambda_1 cosine(F(x_{recon}), F(x)) - \lambda_2 ||x_{recon} - x||_2 - \lambda_3 |a| \tag{13}$$

where $cosine(\cdot, \cdot)$ refers to cosine similarity, $F(\cdot)$ refers to the face recognition model, $\lambda_1, \lambda_2, \lambda_3$ are the weights of losses. The antibody affinity includes three items, the first two items are: the cosine similarity between the denoised facial image and the

original image in the deep feature space; the Euclidean distance between them in the pixel space. These two items measure the difference between the denoised image and the original clean image in both feature space and pixel space, which directly reflects the effectiveness of removing adversarial noise. The third item is the regularization term of the antibody, which constrains the number of eigenvectors contained in the antibody. In other words, it is desired that the antibody can extract the most critical features of the facial image with as few eigenvectors as possible.

Cloning and Mutation:

In the immune system, the cloning and mutation of antibodies are two important mechanisms. On the one hand, antibody cloning can maintain the characteristics of effective antibodies relatively stably. On the other hand, an appropriate degree of mutation during the clonal process allows for antibody variation, enabling the immune system to effectively respond to changes in antigens. It is the effective coordination of these two mechanisms that endow the immune system with strong dynamic adaptability. In the task of adversarial defense for face recognition, it is also necessary to balance the stability and dynamic adaptability of antibodies.

Therefore, in the optimization of antibodies, we achieve the cloning and mutation of antibodies by sampling according to f_e in Eq. 8. As described in Section 3.2, each component of f_e represents the probability of incorporating an eigenvector into the antibody. Therefore, during the optimization process, we sample k antibodies based on f_e, which serves as the cloning process of the antibodies. When the sampling probabilities in f_e are close to 0 or 1, the k antibodies obtained by sampling will have strong consistency, and the probability of antibody mutation is small. Conversely, when the probability within f_e is close to 0.5, the probability of antibody mutation increases, and the differences between the sampled antibodies also become larger. Through the aforementioned sampling process, we complete the cloning and mutation of antibodies.

Antibody Screening and Model Updating:

After obtaining the sampled antibodies $\{a_1, a_2, ..., a_k\}$, the defensive ability of each antibody can be measured using Eq.13: $\{s(a_1), s(a_2), ..., s(a_k)\}$. For antibodies with stronger defensive abilities, we hope to increase their probability of being sampled based on f_e in order to obtain more similar antibodies. Conversely, for antibodies with weaker defensive abilities, we hope to decrease their probability of being sampled, in order to reduce the number of similar antibodies.

To achieve this objective through the updating of model parameters, we first calculate the likelihood of each antibody: $l(a_i)$. The gradient of $l(a_i)$ with respect to the parameters of the proposed model θ represents the direction of increasing likelihood of a_i. Therefore, we adjust the gradient using antibody affinity to obtain the direction of model updating, and then implement model update through gradient descent:

$$\theta \leftarrow \theta - \frac{\phi}{k} \sum_{i=1}^{k} (s_0 - s(a_i)) \nabla_\theta l(a_i), \quad s.t. \quad s_0 = \frac{1}{k} \sum_{i=1}^{k} s(a_i) \tag{14}$$

where ϕ is the learning rate, s_0 provides a baseline for the evaluation of antibodies. The likelihood of antibodies with an affinity higher than s_0 will increase, while the likelihood of antibodies with an affinity lower than s_0 will decrease.

The pseudo-code for the proposed model optimization algorithm is shown in Algorithm 1.

Algorithm 1: Model Optimization Algorithm

Input: raw facial images $\{x^i\}$, corresponding adversarial samples $\{x_{adv}^i\}$, D_θ, number of sampling k, learning rate ϕ

1 for $i \leftarrow 1$ **to** *number of facial images* **do**
2   *get f_e^i by feeding x_{adv}^i into the defense model;*
3   *get $\{a_1^i, a_2^i, ..., a_k^i\}$ by cloning according to f_e^i;*
4   **for** $j \leftarrow 1$ **to** k **do**
5   *get the likelihood $l(a_j^i)$;*
6   *get antibody affinity $s(a_j^i)$ according to Eq.13;*
7   **end**
8   *update θ accroding to Eq.14*
9 end

3.5 Self-supervised Adversarial Training

In order to provide effective guidance for model optimization, we propose a self-supervised adversarial training mechanism. This mechanism involves generating adversarial samples to purposefully train the proposed adversarial defense model. The effectiveness of adversarial training relies on two conditions. Firstly, generating an ample amount of adversarial samples is crucial to prevent the model from taking shortcuts and ensure its generalization. Secondly, the generation process of adversarial samples needs to exhibit consistency to maintain training stability. However, these two aspects are contradictory under the constraint of limited storage space. While ensuring consistency, the limited storage space restricts the number of adversarial samples involved in training.

To address this contradiction, we draw inspiration from MoCo [87] and employ a siamese model $D_{\bar{\theta}}$ for the defense model, where its parameters are updated with an exponential moving average of D_θ:

$$\bar{\theta} \leftarrow \xi\bar{\theta} + (1 - \xi)\theta \tag{15}$$

where $\xi \in (0, 1)$ is the decay rate of updating. As ξ approaches 1, the differences between the adversarial samples from different mini-batches are relatively small, thereby satisfying both the requirement for sample quantity and the demand for consistency. For each facial image x, a corresponding adversarial sample can be generated using this siamese model. When generating adversarial samples, we employ FGSM,

which perturbs the facial sample by making small adjustments in the direction of the gradients of the siamese model:

$$x_{adv} = x + \eta \ sign(\nabla_x L(x)) \tag{16}$$

$$L(x) = 1 - cosine(F(x), F(D_{\bar{\theta}}(x))) \tag{17}$$

where $sign(\cdot)$ refers to sign function, $cosine(\cdot, \cdot)$ refers to cosine similarity, $F(\cdot)$ refers to the face recognition model, η is the scale of adversarial noises.

Through the proposed adversarial self-supervised training, we have established a training mechanism that concurrently fulfills the requirements of both adversarial sample quantity and consistency. This mechanism serves as an effective guidance for the optimization of the adversarial defense model.

3.6 Implementation Details

The implementation details of the proposed adversarial defense method are introduced in this subsection.

The antigen analyzer H in Eq.7 is ResNet-18 [7], which consists of a compendium of 17 convolutional layers complemented by a fully connected layer. The dimension of f_n, which is the output of H, is 512. The number of memory items d_m in Eq.9 is set to 128. The feature mapping layer G in Eq.8 is a fully connected layer, which maps the noise features onto the eigenvector selection space. During the training process, we exclusively consider the top 1500 eigenvectors with the largest eigenvalues and disregard the rest by assigning a probability of zero to their selection. Consequently, the dimension of the eigenvector selection space is 1500, resulting in the dimension of f_e, which is the output of G, is 1500.

The proposed adversarial defense model is trained on the CelebA Dataset [88], which comprises a substantial collection of 202,599 facial images. To ensure consistency, all training images underwent alignment [30] and were resized to 112×112. Additionally, a transformation is applied to convert them to grayscale, facilitating the computation of eigenvectors. The eigenvectors are obtained on the gray-scale facial images of CelebA Dataset.

The decay rate of memory updating ϵ in Eq.12 is set to 0.999. The loss weights $\lambda_1, \lambda_2, \lambda_3$ in Eq.13 are set to 8, 1, and 0.003, respectively. The decay rate of siamese model updating ξ in Eq.15 is set to 0.999. The noise scale during self-supervised adversarial training η in Eq.16 is set to 0.04. The hyper-parameter that determines the number of sampled antibodies, denoted as k in Algorithm 1 is set to a value of 10 during the training. The impact of k will be discussed in Section 4.

Stochastic gradient descent (SGD) with momentum is adopted for training with a batch size of 4. The learning rate is set to 0.01, the momentum is set to 0.9. The face recognition model $F(\cdot)$ in Eq. 13 and Eq. 17 employs ArcFace (ResNet-50) [30]. Prior to conducting self-supervised adversarial training, a warm-up training phase of 50 thousand steps is performed. During this phase, no noise was added to the input images, and the model is solely tasked with completing the facial reconstruction task. Following the warm-up training, the self-supervised adversarial training is carried out for an additional 250 thousand steps.

4 Experiments

In this section, we evaluate the proposed approach through a series of experiments. Firstly, we evaluated the defensive performance of the proposed method by employing the general adversarial attack methods and compared it with the state-of-the-art adversarial defense methods. Following that, we evaluated the defensive performance of the proposed method using adversarial attack techniques tailored for the realm of facial recognition tasks. To conduct a comprehensive examination of the proposed approach, we further employed adaptive attack strategies to rigorously test its efficacy. Moreover, in order to delve deeper into the proposed approach, we conducted an extensive analysis of the generated antibodies by the model. Finally, through ablation studies, we validated the effectiveness of the proposed model as well as the self-supervised adversarial training.

4.1 Evaluation under General Attacking Methods

This subsection employs general adversarial attack methods to evaluate the proposed defense method. We utilize the Equal Error Rate (EER) as the evaluation metric for recognition performance:

$$FRR = \frac{\sum \mathbb{1}(S_p > T)}{N_p} \tag{18}$$

$$FAR = \frac{\sum \mathbb{1}(S_n > T)}{N_n} \tag{19}$$

where FRR is the false rejection rate, FAR is the false acceptance rate, T is the similarity threshold, S_p is the similarities of positive pairs, N_p is the number of positive pairs, S_n is the similarities of negative pairs, N_n is the number of negative pairs. EER refers to the FAR (or FRR) when the threshold T is set in such a way that FAR equals FRR. EER is a concise performance index that serves as a comprehensive evaluation of the discriminative ability of a face recognition model.

Gradient-based adversarial attacks are the most prevalent white-box strategies employed. In the context of white-box attacks, the target model remains fully visible to the attack methods, rendering it an arduous test for defense methods. We have chosen three gradient-based white-box attack methods for testing: FGSM [13] is a classic one-step adversarial attack approach; DeepFool [41] utilizes gradient signals in an iterative manner for adversarial attacking; PGD [43] is the most powerful first-order adversarial attack method. The magnitude of the adversarial noises is quantified by the ratio between the scale of the noise and the scale of the clean image: $I(\zeta) = ||\zeta||/||x||$, where ζ is the adversarial noises. $I(\zeta)$ is set to 0.04 in the experiments of this subsection.

The experiments are conducted using two datasets: Labeled Faces in the Wild (LFW) [89] and MegaFace [90]. For the LFW dataset, we adhere to the official benchmark protocol[1], which involves selecting 3,000 positive pairs and 3,000 negative pairs of images for face verification. Regarding MegaFace, we choose 80 identities from the subset *facescrub* that have more than 50 images per subject. From each identity, we randomly select 10 images for testing. Consequently, there are 7,200 positive pairs and

---

[1]http://vis-www.cs.umass.edu/lfw/pairs.txt

Table 1 Defensive performances on LFW (EER). Lower EER is preferable. The proposed method excels in all three adversarial attacking scenarios while simultaneously delivering comparable performance on clean facial images.

Defense Method	Clean↓	FGSM↓	DeepFool↓	PGD↓
No Defense	0.44%	41.97%	89.49%	99.71%
Quilting [76]	8.77%	9.04%	25.10%	45.79%
TVM [70]	2.95%	19.94%	73.21%	96.62%
PixelDefend [73]	2.05%	18.09%	70.31%	97.70%
MagNet [71]	1.51%	7.86%	14.48%	46.04%
PIN [37]	1.95%	6.15%	7.86%	29.77%
HGD [72]	1.08%	17.35%	20.48%	49.69%
Xie et al. [79]	**0.93%**	20.33%	28.87%	31.29%
MTER [66]	2.62%	10.03%	24.89%	61.06%
Ours	1.01%	**4.46%**	**5.01%**	**14.19%**

Table 2 Defensive performances on MegaFace (EER). Lower EER is preferable. The proposed approach demonstrates superior performance under most adversarial attacking, particularly showing a significant advantage against challenging PGD attacks.

Defense Method	Clean↓	FGSM↓	DeepFool↓	PGD↓
No Defense	1.31%	50.87%	95.41%	99.09%
Quilting [76]	14.23%	18.08%	20.54%	47.13%
TVM [70]	3.66%	21.59%	76.24%	95.12%
PixelDefend [73]	2.41%	25.10%	77.75%	94.03%
MagNet [71]	2.51%	10.79%	24.51%	47.11%
PIN [37]	3.78%	7.82%	9.20%	33.37%
HGD [72]	2.44%	16.15%	29.09%	52.27%
Xie et al. [79]	**1.89%**	16.21%	34.27%	48.76%
MTER [66]	3.08%	11.50%	33.44%	66.95%
Ours	2.23%	**7.27%**	**9.17%**	**18.63%**

632,000 negative pairs available. During testing, adversarial noises are introduced to one image within each pair, while the other image remains unchanged.

We select a set of representative adversarial defense methods for comparison in this subsection. The first group of methods focuses on pixel space denoising, aiming to remove adversarial noises or restore clean images in the pixel space, including Quilting [76], TVM [70], PixelDefend [73], MagNet [71], and PIN [37]. The second group of methods focuses on enhancing the robustness of the recognition model. Among them, MTER [66] solely relies on the adversarial training strategy. Additionally, HGD [72] and Xie et al. [79] combine specially designed model architectures with adversarial training to enhance the robustness of the recognition model. The official implementation of PIN, HGD, and MTER is used in our experiments. PixelDefend, MagNet, and Xie et al. are trained from scratch on the same dataset with the proposed model for a fair comparison. For the denoising-based methods, ArcFace (ResNet-50) [30] is uniformly employed as the recognition model for implementation.

The experimental results on the LFW dataset are presented in Table 1, while the results on the MegaFace dataset are shown in Table 2. The proposed adversarial defense method demonstrates superior defense performance on both datasets in most cases. From the experimental results, it can be observed that denoise-based methods often face a trade-off between the performance on clean samples and adversarial defense performance, making it difficult to strike a balance between the two sides. This is because these methods typically attempt to employ the same denoising approach to counter all adversarial noises, making it challenging to cope with the diversity and complexity of adversarial noises in face recognition. Although the adversarial training based methods achieve satisfactory performance under FGSM attacking, its performance fluctuates significantly under different attacking methods, indicating a lack of generalization to different types of adversarial noises. In contrast, the proposed method is capable of achieving superior defense performance while only sacrificing a slight decrease in recognition performance on clean face images comparing to Xie et al. [79]. This is attributed to the fact that adversarial training methods (including Xie et al. [79]) retrain the facial feature extractor, thereby maintaining its performance on clean images, whereas the proposed method based on noise removal operate under the premise of a fixed facial feature extractor.

4.2 Evaluation under Attacking Methods Tailored for Face Recognition

Due to the importance and uniqueness of facial recognition tasks, researchers have proposed adversarial attack methods specifically targeting facial recognition. In the real-world scenarios, black-box attacks are more common than white-box attacks since the former do not require access to the recognition model's information, making their execution plainer and more straightforward. To ascertain the effectiveness of the proposed method against black-box attacks, we experimented with various forms of black-box attacks tailored for face recognition in this subsection. Compared to general adversarial attack methods, these methods pose a greater threat to the practical application of facial recognition systems.

The adversarial attack methods employed include the following three: *DFANet* [40]: DFANet is a transfer-based black-box attack method specifically designed for facial recognition models. It adopts gradient-based attack techniques for facial recognition tasks, making it a powerful transfer-based attacking method for face recognition. *Evolutionary* [34]: Evolutionary is a black-box attack method targeting facial recognition systems. It directly exploits the decision outcomes of the facial recognition system, without requiring access to the gradient information of the recognition model. *Sticker attacking* [35, 56]: Sticker attacking is a category of black-box attack methods that operate in the physical domain. By making subtle modifications to the physical appearance of the target object, it produces powerful attacks and can be easily carried out by an attacker without any prior knowledge about the recognition model. These methods significantly reduce the cost of adversarial attacks, posing a substantial threat to facial recognition systems. The comparative methods employed in this subsection remain consistent with the ones used in Section 4.1. Similarly, ArcFace is employed as the recognition model for conducting experiments on denoising-based methods as well.

Table 3 The experimental results on TALFW. The proposed method demonstrates the most outstanding defensive performance.

Defense Method	EER↓
No Defense	39.71%
Quilting [76]	24.33%
TVM [70]	30.90%
PixelDefend [73]	33.22%
MagNet [71]	22.50%
PIN [37]	21.95%
HGD [72]	39.23%
Xie et al. [79]	40.33%
MTER [66]	39.16%
Ours	**20.43%**

DFANet:

The experiments under the DFANet attacking are conducted using the TALFW dataset [91]. The adversarial samples of TALFW are crafted by utilizing DFANet, with the source images being obtained from LFW. TALFW serves as an official test benchmark offered by DFANet. The evaluation on TALFW still utilizes EER as the performance index for assessment.

The testing results on TALFW are presented in Table 3. The proposed method demonstrates superior defensive performance compared to the comparative methods. Through experimentation, it can be observed that methods based on adversarial training perform almost on par with the performance without any defense mechanisms, exhibiting a significant gap when compared to methods based on denoising. Adversarial training based methods extensively incorporate adversarial samples into the training set and continually seek out vulnerabilities in the recognition model. However, since adversarial training is a dynamic process with new data constantly emerging, the model often focuses only on the latest training data during the training process, which can lead to a situation where fixing one problem leads to another. This results in the performance of the model being nearly identical to that of models without any defense mechanisms when facing attacking of DFANet. On the other hand, the proposed memory module possesses the capacity to preserve adversarial noise patterns, thereby mitigating the challenges associated with adversarial training throughout the self-supervised adversarial training procedure.

Evolutionary:

When employing the Evolutionary method to adversarial attacks, the attacking process revolves around continuously optimizing the adversarial samples under the constraint of achieving successful attacks. The objective is to gradually minimize the disparity between the adversarial samples and the clean facial images. Hence, we utilize the defense performance metric proposed officially within the realm of Evolutionary methods, namely the mean squared error (MSE) between the adversarial samples and the clean facial images under the same number of optimization steps.

18

Table 4 MSE on LFW under Evolutionary attacking. At the same step of the iteration, a higher average distortion indicates better performance of the defense method. The proposed method surpasses the listed comparative methods.

Number of Attack Steps		1000 ↑	5000↑	10000↑
	No Defense	3.1e-3	2.5e-4	8.9e-5
	MagNet [71]	1.0e-2	7.7e-3	5.8e-3
	PIN [37]	7.5e-2	5.5e-2	3.1e-2
Dodging	HGD [72]	8.7e-3	6.3e-3	3.9e-3
	Xie et al. [79]	7.4e-3	5.8e-3	2.7e-3
	MTER [66]	9.9e-3	6.5e-3	4.9e-3
	Ours	**1.1e-1**	**7.5e-2**	**4.7e-2**
	No Defense	2.4e-3	2.4e-4	5.9e-5
	MagNet [71]	8.3e-3	5.7e-3	3.1e-3
	PIN [37]	4.6e-2	3.6e-2	2.9e-2
Impersonation	HGD [72]	1.1e-3	7.2e-4	4.8e-4
	Xie et al. [79]	9.7e-4	6.6e-4	4.1e-4
	MTER [66]	1.7e-3	9.3e-4	6.1e-4
	Ours	**1.3e-1**	**9.2e-2**	**7.9e-2**

The larger this error, the more challenging the attack becomes, signifying a stronger defense performance. There are two experimental configurations: dodging and impersonation. Dodging refers to adversarial attacks attempting to recognize positive facial image pairs as negative ones, while impersonation is the opposite, attempting to recognize negative facial image pairs as positive ones. The experimental testing dataset employed in this experiment remains consistent with those in Section 4.1. Due to their high computational complexity, Quilting, TVM, and PixelDefend are difficult to implement for Evolutionary attacking experiments. Consequently, we select the remaining comparative methods discussed in Section 4.1 for this experimentation.

The experimental results are shown in Table 4 and Table 5. The experimental results demonstrate that the proposed approach outperforms other methods in most cases across both datasets. As Evolutionary attack relies on continuously exploring the decision boundaries of the target recognition model to optimize adversarial noises, the performance metrics at larger numbers of attack iterations better reflect the effectiveness of defense methods. The advantage of the proposed approach becomes more prominent with a larger number of attack steps. This is attributed to the fact that the proposed adversarial defense methodology exhibits the capacity to analyze and retain crucial information from each facial sample, leading to a more robust decision boundary. As a corollary, the task of mounting the Evolutionary attack becomes noticeably more formidable.

Sticker Attacking:

Sticker attacking is a category of physical domain adversarial attack methods, where adversarial stickers are applied to specific areas of faces to introduce adversarial noises during the image acquisition process. Adversarial stickers can be directly applied to the face [56] or attached to accessories [92], and can even be achieved through makeup

Table 5 MSE on MegaFace under Evolutionary attacking. It particularly demonstrates the advantages of the proposed method when the number of iterations increases, as it provides a better reflection of the performance of defense methods.

Number of Attack Steps		1000↑	5000↑	10000↑
Dodging	No Defense	3.5e-3	8.5e-4	9.7e-5
	MagNet [71]	**9.8e-2**	6.5e-3	2.2e-3
	PIN [37]	9.2e-2	**8.4e-2**	6.7e-2
	HGD [72]	8.8e-2	6.0e-3	1.9e-3
	Xie et al. [79]	7.2e-3	5.5e-3	2.3e-3
	MTER [66]	9.7e-3	6.6e-3	5.1e-3
	Ours	9.5e-2	8.3e-2	**6.9e-2**
Impersonation	No Defense	2.4e-3	1.1e-4	5.5e-5
	MagNet [71]	7.9e-3	2.1e-3	1.4e-3
	PIN [37]	**7.7e-2**	6.8e-2	6.2e-2
	HGD [72]	7.2e-3	1.7e-3	8.4e-4
	Xie et al. [79]	6.5e-3	3.7e-3	1.8e-3
	MTER [66]	7.9e-3	4.6e-3	2.9e-3
	Ours	7.6e-2	**7.1e-2**	**6.6e-2**

techniques [93]. These methods do not require intervention in the data processing pipeline of the face recognition system, making them cost-effective and posing a threat to the actual deployment of face recognition systems.

In practical scenarios, both dodging and impersonating types of sticker attacks are prevalent. But these two forms of attack typically target different modules of face recognition systems. Dodging attacks usually deceive the face detection module to prevent the attacker from being detected, while impersonating attacks, conducted after a face has been successfully detected, usually confuse the facial feature extraction component to misidentify the attacker as another individual. Actually, simply obscuring key regions of the face (e.g., by wearing masks or hats) suffices to achieve the desired result of dodging attacks. If faces can not be successfully detected, adversarial defenses for facial feature extraction are not needed. Given that this work focuses on adversarial defenses in facial feature extraction, we have specifically conducted experiments on impersonating attacks.

In our experiments, we employ AdvHat [92] as the adversarial attack method. We simulated the generation process of adversarial stickers and then applied the generated stickers to the facial images for the attack, as shown in Fig. 3. Unlike other adversarial attack methods, the sticker attack does not impose a restriction on the intensity of adversarial noises.

A total of 1000 facial images from different individuals are randomly selected from the LFW for the test. The generated adversarial sticker is applied on the foreheads of these test facial images. The objective of the adversarial attack is to manipulate the facial recognition model into identifying all 1000 test faces as the target identity. The size of the adversarial sticker was set to 20×72. The decision similarity threshold for facial recognition was deliberately set at 0.2 to ensure that the recognition model (ArcFace) correctly identifies the majority of positive pairs while maintaining
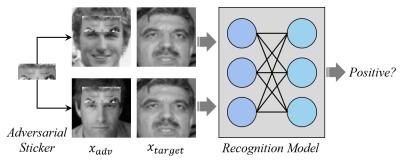
Fig. 3 The sticker attacking employs the adversarial sticker strategically overlaid on specific regions of facial images to induce mistakes in recognition models.

Table 6 The accuracy of recognition under sticker attacking. The proposed method surpassed all comparative methods, yielding the best results.

Defense Method	Accuracy ↑
No Defense	25.6%
Quilting [76]	92.2%
TVM [70]	86.0%
PixelDefend [73]	92.7%
MagNet [71]	95.9%
PIN [37]	97.6%
HGD [72]	89.1%
Xie et al. [79]	90.7%
MTER [66]	93.4%
Ours	**98.4%**

an acceptable false negative rate (TAR=99.89%, FRR=2.09%) on clean facial images. This stringent configuration serves as a rigorous benchmark for evaluating defense methods. The accuracy of recognition serves as the performance index in the experimentation. When the similarity between a test facial image with the added adversarial sticker and the target facial image is below 0.2, it is considered a correct identification. Conversely, if the similarity exceeds 0.2, it is considered an incorrect identification. A higher accuracy indicates better performance of the defense method.

The recognition accuracy under sticker attacking is shown in Table 6. Experimental results demonstrate that the proposed method achieved the best recognition outcomes. The adversarial noises in the physical domain differ significantly from that in the digital domain, as their intensity is no longer constrained, posing a great challenge for noise removal. Due to the significant alterations caused by adversarial stickers, the resulting facial images deviate notably from the distribution of clean facial images. Therefore, the key to denoising in such cases lies in effectively modeling and utilizing the distribution of clean facial images. The proposed method achieves this through the designed antibodies. Moreover, the proposed adversarial defense approach offers an ability to provide customized defensive measures for each individual sample, making it arduous for adversarial stickers to have a universal impact across diverse facial

Table 7 The experimental results of the adaptive attacking (EER).

Attack Method	FGSM↓	DeepFool↓	PGD↓
No Defense	16.39%	33.55%	54.86%
Ours	10.76%	24.92%	38.68%

samples. As a result, the practical applicability of sticker attacks in real-world scenarios is significantly diminished.

4.3 Evaluation under Adaptive Attacking

For real-world deployment of face recognition systems, if attackers become aware of adversarial defense mechanisms, they are likely to face adaptive attacking. An adaptive attack poses a formidable challenge. In this type of attack, attackers not only target the recognition model but also optimize their strategies to circumvent the defense mechanisms. These attacks rigorously test the efficacy of defense methods, as attackers have the capability to adjust their tactics based on the employed defenses, leaving the defense mechanisms unable to adapt in response.

In this subsection, we present an adaptive attack against the proposed adversarial defense method to evaluate its defensive capabilities. Given that the proposed defense method is built upon antibodies, the composition of antibodies serves as the foundation of the entire method. Therefore, in the adaptive attack, all possible antibodies involved in the proposed method are considered within the attack scope. Specifically, during the optimization of adversarial noises, the input facial images undergo preprocessing with a specific antibody a^*, which consists of the top 1500 eigenvectors with the largest eigenvalues:

$$a^* = \{e_1, e_2, ..., e_{1500}\} \tag{20}$$

This means that all possible antibodies are subsets of a^*, which means the optimization direction will encourage the adversarial noises to avoid being removed by any possible antibodies. This poses a rigorous test for the proposed method. We utilized the three adversarial attack methods outlined in Section 4.1 for the optimization of adversarial noises in the adaptive attack. The intensity of the adversarial noises $I(\zeta)$ is also set as 0.04, and the experiment is conducted on LFW.

The experimental results are presented in Table 7. Comparing with Table 1, the experimental results reveal a significant decline in recognition performance of the proposed defense method when subjected to adversarial noise generated by the adaptive attacking. This can be attributed to the targeted nature of the adaptive attacking, which diminishes the effectiveness of the antibodies generated by the defense method. However, even under the adaptive attacking, the proposed method still outperforms most of the comparative methods listed in Table 1. Furthermore, when contrasting with the first row of Table 1, it is evident that the adaptive attacking have a significantly reduced impact on the model without any adversarial defense measures. This indicates the substantial cost incurred by adversarial noises in bypassing the proposed defense method, resulting in a greatly weakened attack effect in the absence of any defense measures. This further highlights the superior defensive performance of the proposed method.
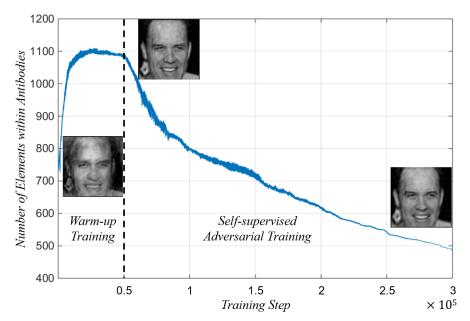
22

Fig. 4 The number of eigenvectors contained within antibodies during the training process. During the training process, the number of eigenvectors present in the antibodies initially rises and then declines. This progression enables the antibodies to first enhance their reconstruction capabilities, followed by a selective refinement of eigenvectors to bolster their denoising prowess. As a result, the antibodies achieve the remarkable ability to effectively eliminate adversarial noise while preserving vital facial features.

4.4 Quantitative Analysis of Antibodies

In order to conduct a more comprehensive analysis of the proposed defense method, this section focuses on the quantitative analysis of the antibodies generated by the proposed method, unveiling the evolutionary process of antibodies during model training. This analysis primarily encompasses three aspects of antibodies: their sparsity, mutation probability, and specificity.

Sparsity of Antibodies:

The sparsity of antibodies is determined by the number of eigenvectors they contain, and a smaller quantity of eigenvectors indicates a higher level of sparsity in the antibodies. The number of eigenvectors present in antibodies represents their selection of features for noise removal and facial sample reconstruction. A greater number of eigenvectors suggests that the antibodies are biased towards reconstructing more intricate details of the input samples, while a smaller number indicates their inclination towards extracting a smaller subset of essential features. To quantitatively measure the sparsity of antibodies, we employ the number of eigenvectors contained within them as a metric to evaluate their sparsity.

The number of eigenvectors contained within antibodies during the entire training process is depicted Fig. 4 (average number of antibodies per mini-batch). Several key observations can be summarized from Fig. 4:

23

- During the initial warm-up training phase, the number of eigenvectors in antibodies rapidly increases. This is because, at this stage, the model focuses solely on facial image reconstruction without incorporating self-supervised adversarial training. By increasing the number of eigenvectors in antibodies, the model can reduce reconstruction errors. Therefore, during antibody optimization, the model quickly increases the number of eigenvectors, thereby reducing antibody sparsity and improving reconstruction results.
- In the warm-up training phase, the number of eigenvectors in antibodies reaches around 1100 and fluctuates around this value. This occurs due to the regularization term in Eq.13, which penalizes the number of elements in antibodies. The benefits of adopting more eigenvectors are suppressed by the penalty imposed by the regularization term in Eq.13.
- Once self-supervised adversarial training is introduced, the number of eigenvectors in antibodies gradually decreases, indicating an increase in antibody sparsity. This is because, in self-supervised adversarial training, the model no longer focuses solely on reconstructing input samples but tracks its target at removing adversarial noise by refining the eigenvectors.
- The Fig.5 illustrates the recovery effects of antibodies at different training stages. When training concludes, the number of eigenvectors in generated antibodies has decreased to around 500. Although this number is lower compared to the initial 750, the recovery effects have significantly improved compared to the initial phase of training. Compared to the end of the warm-up training phase, the antibodies at the final stage retain crucial facial information and effectively remove adversarial noise by filtering out some eigenvectors.
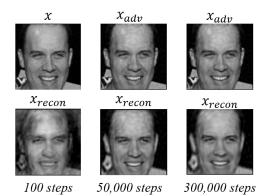

Fig. 5 Examples of the recovery performance of antibodies on input face images at different stages of training.

Through the aforementioned analysis, it can be observed that the number of eigenvectors contained within the antibodies initially increases and then decreases during the entire training process. However, during this process, the defense model does not stand still but goes through a process of first improving its reconstruction ability and then refining the denoising ability by eliminating crude components from the

eigenvectors. Eventually, this results in an effective removal of adversarial noise while preserving the key features of the face.

Mutation Probability:

During the optimization process of the proposed defense model, the antibody mutation plays a crucial role. Adequate antibody mutation during training not only enhances antibody diversity but also effectively prevents the model from getting trapped in local optima during optimization. However, if the degree of antibody mutation is excessively high, it may lead to difficulties in the convergence of the defense model. In order to further investigate the optimization process of the proposed defense method, we analyze the probability of antibody mutation.

The antibody mutation refers to the changes that occur in the eigenvectors comprising the antibodies. In order to quantitatively measure the probability of antibody mutation, we calculate the average probability of reversing the selection of antibodies for each eigenvector (whether to include it or not) as a metric of evaluation:

$$P_{mutation} = \frac{1}{k} \sum_i^k (0.5 - |(f_e(i) - 0.5)|) \tag{21}$$

where k is the dimension of f_e, $|\cdot|$ refers to absolute value. The closer $f_e(i)$ is to 0.5, the more likely it is for the selection of the i-th eigenvector to undergo mutation.

The mutation probability of antibodies during the entire training process is shown in Fig. 6 (the mean of each mini-batch). From Fig. 6, we can draw several conclusions:

- During the initial state, the probability of mutation for antibodies is high, but during the warm-up training phase, this probability decreases rapidly. This is due to the model quickly honing in on the critical eigenvectors for face image reconstruction, resulting in a rapid decrease in the mutation probability of these eigenvectors.
- Upon entering the self-supervised adversarial training phase, the mutation probability of antibodies slightly increases. This is due to a change in the training objective of the model, and the model adapts by increasing the mutation probability of antibodies to better fit this new objective.
- The mutation probability of antibodies eventually converges and approaches 0, indicating that the model gradually finds the best antibody for each input sample after training, and no longer mutates.

Through the fluctuations in the mutation probability of antibodies, we can observe how the proposed defense method possesses the capability to adaptively modify the mutation probability to align with the evolving training objective. As the training objective undergoes changes, the model augments the mutation probability, promoting the flexibility of the defense system and diversifying the antibodies. Consequently, this enables a readaptation to the new objective at hand. Conversely, when the training objective stabilizes, the model progressively diminishes the mutation probability of antibodies, fostering convergence. Thus, the proposed method exhibits dynamic adaptability akin to that of an immune system.
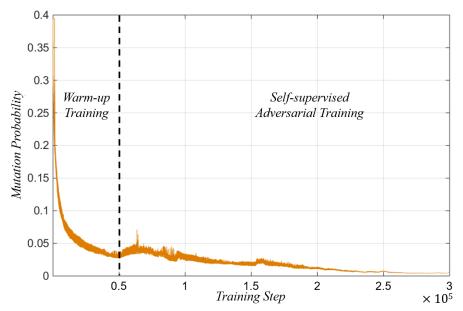
Fig. 6 The mutation probability of antibodies during the training process. The proposed defense approach demonstrates the ability to adaptively adjust the mutation probability of antibodies during the training process, aligning them with the evolving training objective. When the training objective undergoes changes, the model responds by increasing the mutation probability, thereby enhancing the overall flexibility of the defense model. Conversely, as the training objective stabilizes, the model gradually reduces the mutation probability of antibodies, allowing for a progressive convergence of the model.

Specificity of Antibody:

Due to the particular nature of facial recognition tasks, the specificity and diversity of adversarial noises in facial recognition are remarkably pronounced. Therefore, we aspire for the defense approach to furnish tailored noise removal strategies for each input facial image, thereby addressing this formidable challenge. In order to investigate whether the proposed defense method indeed confers specificity to the antibodies, we examine the specificity of antibodies during the training process.

The specificity of antibodies can be discerned through differences manifested between antibodies of different samples. To quantitatively assess the specificity of antibodies, we propose a metric that measures the dissimilarity between two antibodies:

$$J(a_i, a_j) = \#\{e | (e \in a_i \wedge e \notin a_j) \vee (e \notin a_i \wedge e \in a_j)\} \tag{22}$$

which means the number of eigenvectors that are uniquely present in either one of the antibodies. Based on $J(a_i, a_j)$ it is possible to quantitatively measure the specificity of a set of antibodies:

$$V(\{a_1, a_2, ..., a_n\}) = \frac{1}{n(n-1)} \sum_{i}^{n} \sum_{j \neq i}^{n} J(a_i, a_j) \tag{23}$$
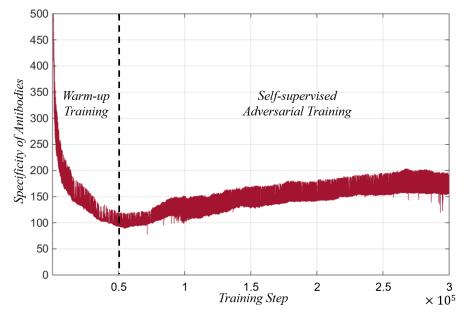
26

Fig. 7 The specificity of antibodies during the training process. The specificity of antibodies undergoes a process of initial decline followed by a subsequent recovery during the training. The decrease in specificity during the warm-up training phase can be primarily attributed to the rapid reduction in antibody mutation probability. On the other hand, the gradual increase in antibody specificity observed during the self-supervised adversarial training is a result of the model progressively providing tailored noise removal strategies for each individual sample to effectively eliminate adversarial noise.

The specificity of each mini-batch of antibodies throughout the entire training process is illustrated Fig. 7. Several observations can be made from Fig. 7:

- At the beginning of the training process, the specificity of the antibodies is initially high, due to the high probability of antibody mutation (as shown in the Fig. 6) and the instability of the antibodies. After the warm-up training, the specificity rapidly decreases to 100. The reason is that facial images contain shared characteristics and the goal of the warm-up training is to reconstruct facial images, rendering high antibody specificity unnecessary.
- Upon commencing self-supervised adversarial training, the specificity of the antibodies undergoes a gradual enhancement and eventually settles at approximately 200. This implies that, on average, there are 200 distinct eigenvectors between every pair of antibodies. It indicates that the defense model progressively commences providing noise removal manners tailored to each specific adversarial sample during the course of self-supervised adversarial training.

Although the specificity of the antibodies experienced a process of decline and then recovery during the training process, this does not mean that the defense model returned to the initial state. The decline in specificity observed during the warm-up training phase primarily stems from the swift reduction in antibody mutation probability and the enhancement of antibody stability. Conversely, the gradual augmentation

27

Table 8 The experimental results of ablation study for self-supervised adversarial training and memory module (EER). SSAT refers to self-supervised adversarial training. These results also demonstrate the significant improvement in the defense capability of the model by self-supervised adversarial training and memory module.

Attack Method	Clean↓	FGSM↓	DeepFool ↓	PGD↓
w/o SSAT	1.04%	11.99%	12.08%	40.60%
w/o Memory	1.06%	9.06%	9.75%	22.41%
Ours	**1.01%**	**4.46%**	**5.01%**	**14.19%**

of antibody specificity throughout the self-supervised adversarial training arises from the model's provision of tailored noise removal solutions for each individual sample, thereby eliminating adversarial noises.

4.5 Ablation Study

In this subsection, we first verify through experiments the impact of self-supervised adversarial training and the memory module on defense performance, then explore the applicability of the proposed defense method to different recognition models. Finally, we conduct sensitivity analysis on key hyper-parameters in the model optimization process.

Ablation Study for Self-supervised Adversarial Training:

Self-supervised adversarial training is a crucial component of the proposed defense method. Its purpose is to enhance the noise removal ability of defense models by introducing adversarial noises during the training process. To evaluate the effectiveness of this training strategy, we compared the model without self-supervised adversarial training to the original model. We conducted experiments on LFW using three types of adversarial attacks: FGSM, DeepFool, and PGD. The experimental protocols are consistent with Section 4.1.

The experimental results are shown in Table 8. The results indicate that the impact of self-supervised adversarial training on the recognition of clean facial images is not significant. However, the performance of the model without self-supervised adversarial training exhibited a significant decline in adversarial defense, especially under the strongest PGD attack. This is because, if self-supervised adversarial training is not conducted, the model's training objective is face image reconstruction, and it will tend to incorporate all eigenvectors into the antibody to reconstruct all the details of the input image. Although the regularization term in Eq. 13 constrains the number of eigenvectors contained in the antibody to encourage the model to focus on important eigenvectors, the model's performance in adversarial defense is significantly reduced because it has not undergone targeted adversarial noise removal training. These results also demonstrate the significant improvement in the defense capability of the model by self-supervised adversarial training.

Table 9 Experimental results in conjunction with MobiFace (EER).

Attack Method	Clean↓	FGSM↓	DeepFool ↓	PGD↓
MobiFace w/o defense	0.67%	57.72%	82.58%	99.19%
MobiFace with defense	1.56%	7.37%	7.59%	24.82%
ArcFace w/o defense	**0.44%**	41.97%	89.49%	99.71%
ArcFace with defense	1.01%	**4.46%**	**5.01%**	**14.19%**

Ablation Study for Memory Module:

Drawing inspiration from the memory mechanisms of the immune system, a memory module is incorporated into the proposed defense model. The memory module can store noise patterns during the training process and guide for generating antibodies through memory retrieval. To verify the effectiveness of the memory module, we trained and tested the model without using the memory model and compared its defense performance with the original model. When the memory module is not utilized, $\hat{f}_n$ in Eq. 8 is simply substituted with f_n. We conducted experiments using three adversarial attack methods, namely FGSM, DeepFool, and PGD, on the LFW dataset. The testing protocol of the experiment is also consistent with Section 4.1.

The experimental results are shown in Table 8. The results indicate that when the memory module is not utilized, the proposed method performs similarly to the original model on clean facial images, but the adversarial defense performance decreases. This is because the recognition performance on clean facial images only depends on the model's reconstruction ability, while the adversarial defense capability also requires the model to have strong noise removal ability. The noise features stored in the memory module can provide guidance for generating antibodies, preventing the model from the scenario where fixing one vulnerability leads to the emergence of another.

Transferability to Different Face Recognition Models:

In previous experiments, face recognition models that were compatible with the proposed defense method all utilized ArcFace (ResNet-50) [30]. The purpose of conducting this experiment is to answer the question of whether the proposed defense method can be implemented with other face recognition models for adversarial defense. MobiFace [39], as a lightweight face recognition model, has many differences in design and structure compared to ArcFace. MobiFace is utilized in tandem with the proposed defense model without undergoing retraining, but instead, directly integrated with it. The experiments also employed FGSM, DeepFool, and PGD as adversarial attack methods, conducted on LFW with experimental settings consistent with Section 4.1.

The experimental results are shown in Table 9. It can be observed that the proposed defense model performs worse than ArcFace when used in conjunction with MobiFace. This is mainly due to two reasons: Firstly, MobiFace itself is a lightweight model with weaker recognition ability than ArcFace, which can be demonstrated by the performance difference between them on clean facial images. Secondly, the cooperation between ArcFace and the defense model is more harmonious, as ArcFace participates as the face recognition model in both antibody affinity measurement and adversarial self-supervised training. However, even so, the proposed method still outperforms the

Table 10 Sensitivity analysis on the number of cloned antibodies k.

Attack Method	Clean↓	FGSM↓	DeepFool ↓	PGD↓
$k = 10$	1.01%	4.46%	5.01%	14.19%
$k = 20$	1.12%	4.43%	5.13%	14.09%
$k = 30$	**0.87%**	**4.36%**	**4.85%**	**13.98%**

compared methods in Table 1, which indicates its ability to be directly applied to other face recognition models without retraining.

Sensitivity Analysis on Hyper-parameters:

There is a key hyper-parameter in Algorithm 1, which is the number of cloned antibodies k at each iteration. To investigate the effect of the number of cloned antibodies on the defense performance, we conducted comparative experiments with different values of k, namely $k = 10$, $k = 20$, and $k = 30$. Other than the number of cloned antibodies, the training settings for the three defense models were consistent with Section 3.6. FGSM, DeepFool, and PGD are utilized as adversarial attack methods for testing on LFW, with experimental settings consistent with Section 4.1.

The experimental results are shown in Table 10. When $k = 10$ and $k = 20$, the model's performance is quite comparable. When $k = 30$, the model's performance is optimal, in terms of both recognition accuracy on clean facial images and its ability to defend against adversarial attacks. In general, increasing the number of cloned antibodies can enhance the model's defense performance. This is due to the fact that increasing the quantity of cloned antibodies enables the model to undertake a broader exploration during the optimization process, and also facilitates the exploitation of antibody mutation. It is worth noting, however, that the performance improvement resulting from increasing the number of cloned antibodies is limited, and it also amplifies the computational complexity of model optimization.

4.6 Failure Cases Analysis

In this subsection, we delve into the analysis of the model's failure cases. Such an examination is instrumental in gaining a more profound insight into the proposed model, significantly enhancing our understanding of its defensive performance.

Fig. 8 Examples of failure cases. The first three samples of defense failure exhibit significant head postures. The last sample is an occluded face image.

Fig. 8 displays four samples where adversarial defense failed. It is evident from Fig. 8 that the first three samples of defense failure exhibit significant head postures.

The reason for the model's failure on these samples lies in the composition of the antibodies. The facial images with large head postures exhibit a substantial difference in data distribution compared to frontal face images. Since the eigenvectors constituting the antibodies are derived from a data distribution predominantly composed of frontal face data (CelebA), their capacity to recover data with significant head postures, such as side faces, is limited. This limitation adversely affects the defensive capabilities of the antibodies.

The fourth failure sample is an occluded face image. Occlusions have a notable impact on face feature extraction, and the intra-class distance between clean images significantly increases due to occlusions. This leads to samples, post adversarial noise removal, being challenging to recognize normally.

5 Conclusion

In response to the challenges posed by the specificity and diversity of adversarial noises in face recognition, we draw inspiration from the working mechanism of the immune system and propose an adversarial defense method specifically designed for face recognition tasks in this paper. Extensive experimental results demonstrate the efficacy of the proposed method, surpassing state-of-the-art adversarial defense methods. Through a series of experiments and analyses, the advantages of the proposed method can be summarized as follows:

- Regarding the strong specificity of adversarial noise in facial recognition, the proposed method can offer specific noise removal strategies for each input sample. This enables the effective removal of adversarial noises while preserving essential facial features.
- Through the proposed self-supervised adversarial training, the contradiction between the quantity and consistency of adversarial samples is resolved, thereby providing effective guidance for the optimization of defense models.
- The proposed method possesses dynamic adaptability, allowing it to autonomously adjust the optimization process of the defense model to accommodate changes in training data and objectives. This aspect holds inspiring implications for researchers to design novel methodologies based on this foundation.
- The proposed method demonstrates a high level of applicability across various facial recognition models, as it can be directly employed without the need for retraining when applied to different facial recognition systems.
- The Artificial Immune System proposed in this paper also offers inspiration for other face-related security tasks, such as defense against face presentation attacks and DeepFakes attacks. Incorporating the principles of antibody cloning, mutation, selection, and memory mechanisms into these tasks could enhance the model's performance, particularly in terms of dynamic adaptability and generalization capabilities against various attack methodologies. This improvement presupposes the design of appropriate antibody forms and optimization objectives.

Data Avaliability Statement

The data that support the findings of this study are available in Large-scale CelebFaces Attributes (CelebA) Dataset: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, Labeled Faces in the Wild: http://vis-www.cs.umass.edu/lfw/, and MegaFace Dataset: http://megaface.cs.washington.edu/dataset/download.html.

Acknowledgement

References

[1] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)

[3] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations, pp. 1–10 (2015)

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

[5] Huang, G., Liu, Z., Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[6] Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[7] 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)

[8] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28**, 91–99 (2015)

[9] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

[10] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

[11] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)

[12] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (2013)

[13] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

[14] Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)

[15] Wei, X., Yu, J., Huang, Y.: Infrared adversarial patches with learnable shapes and locations in the physical world. International Journal of Computer Vision, 1–17 (2023)

[16] Zheng, Z., Zheng, L., Yang, Y., Wu, F.: U-turn: Crafting adversarial queries with opposite-direction features. International Journal of Computer Vision **131**(4), 835–854 (2023)

[17] Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. In: Proceedings of the International Conference on Learning Representations (2017)

[18] Taesik, N., Hwan, K.J., Saibal, M.: Cascade adversarial machine learning regularized with a unified embedding. In: Proceedings of the International Conference on Learning Representations (2018)

[19] Florian, T., Alexey, K., Nicolas, P., Ian, G., Dan, B., Patrick, M.: Ensemble adversarial training: Attacks and defenses. In: Proceedings of the International Conference on Learning Representations (2018)

[20] Ross, A.S., Doshi-Velez, F.: Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and

Thirtieth Innovative Applications of Artificial Intelligence Conference (2018)

[21] Shao, R., Perera, P., Yuen, P.C., Patel, V.M.: Open-set adversarial defense with clean-adversarial mutual learning. International Journal of Computer Vision **130**(4), 1070–1087 (2022)

[22] Dolatabadi, H.M., Erfani, S.M., Leckie, C.: Adversarial coreset selection for efficient robust training. International Journal of Computer Vision **131**(12), 3307–3331 (2023)

[23] Zhang, Y., Hou, J., Yuan, Y.: A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks. International Journal of Computer Vision, 1–33 (2023)

[24] Liu, A., Tang, S., Liu, X., Chen, X., Huang, L., Tu, Z., Song, D., Tao, D.: Towards defending multiple adversarial perturbations via gated batch normalization. International Journal of Computer Vision (2023)

[25] Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2014)

[26] Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2014)

[27] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2015)

[28] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 212–220 (2017)

[29] Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5265–5274 (2018)

[30] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698 (2018)

[31] Ren, M., Wang, Y., Sun, Z., Tan, T.: Dynamic graph representation for occlusion handling in biometrics. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11940–11947 (2020)

[32] Min, R., Yunlong, W., Yuhao, Z., Kunbo, Z., Zhenan, S.: Multiscale dynamic

graph representation for biometric recognition with occlusions. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 15120–15136 (2023)

[33] Yuhao, Z., Min, R., Hui, J., Linlin, D., Zhenan, S., Ping, L.: Joint holistic and masked face recognition. IEEE Transactions on Information Forensics and Security **18**, 3388–3400 (2023)

[34] Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J.: Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7714–7722 (2019)

[35] Komkov, S., Petiushko, A.: Advhat: Real-world adversarial attack on arcface face id system. In: Proceedings of the International Conference on Pattern Recognition (2021)

[36] M., B.F.: A modification of jerne's theory of antibody production using the concept of clonal selection. The Australian journal of science **20**, 67–69 (1957)

[37] Min, R., Yuhao, Z., Yunlong, W., Zhenan, S.: Perturbation inactivation based adversarial defense for face recognition. IEEE Transactions on Information Forensics and Security **17**, 2947–2962 (2022)

[38] Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security, 2884–2896 (2018)

[39] Duong, C.N., Quach, K.G., Jalata, I., Le, N., Luu, K.: Mobiface: A lightweight deep learning face recognition on mobile devices. In: IEEE 10th International Conference on Biometrics Theory, Applications and Systems, pp. 1–6 (2019). IEEE

[40] Zhong, Y., Deng, W.: Towards transferable adversarial attack against deep face recognition. IEEE Transactions on Information Forensics and Security **16**, 1452–1466 (2020)

[41] Seyed-Mohsen, M.-D., Alhussein, F., Pascal, F.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)

[42] Nicholas, C., David, W.: Towards evaluating the robustness of neural networks. In: 2017 Ieee Symposium on Security and Privacy (sp), pp. 39–57 (2017). Ieee

[43] Aleksander, M., Aleksandar, M., Ludwig, S., Dimitris, T., Adrian, V.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Learning Representations (2018)

[44] Jiawei, S., Vasconcellos, V.D., Kouichi, S.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019)

[45] Zhang, J., Huang, J.-t., Wang, W., Li, Y., Wu, W., Wang, X., Su, Y., Lyu, M.R.: Improving the transferability of adversarial samples by path-augmented method. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8173–8182 (2023)

[46] Zhang, J., Huang, Y., Wu, W., Lyu, M.R.: Transferable adversarial attacks on vision transformers with token gradient regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16415–16424 (2023)

[47] Liu, Z., Xu, Y., Ji, X., Chan, A.B.: Twins: A fine-tuning framework for improved transferability of adversarial robustness and generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16436–16446 (2023)

[48] Wang, Z., Yang, H., Feng, Y., Sun, P., Guo, H., Zhang, Z., Ren, K.: Towards transferable targeted adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20534–20543 (2023)

[49] Liang, K., Xiao, B.: Styless: Boosting the transferability of adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8163–8172 (2023)

[50] Chaoning, Z., Philipp, B., Adil, K., So, K.I.: Data-free universal adversarial perturbation and black-box attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7868–7877 (2021)

[51] Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., Ren, K.: Feature importance-aware transferable adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7639–7648 (2021)

[52] Yuan, Z., Zhang, J., Jia, Y., Tan, C., Xue, T., Shan, S.: Meta gradient adversarial attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7748–7757 (2021)

[53] Li, Z., Yin, B., Yao, T., Guo, J., Ding, S., Chen, S., Liu, C.: Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24626–24637 (2023)

[54] Qian, L., Yuxiao, H., Ye, L., Dongxiao, Z., Xin, J., Yuntian, C.: Discrete pointwise attack is not enough: Generalized manifold adversarial attack for face

recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20575–20584 (2023)

[55] Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 Acm Sigsac Conference on Computer and Communications Security, pp. 1528–1540 (2016)

[56] Yang, X., Liu, C., Xu, L., Wang, Y., Dong, Y., Chen, N., Su, H., Zhu, J.: Towards effective adversarial textured 3d meshes on physical face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4119–4128 (2023)

[57] Lei, H., Yun-Yun, T., Pin-Yu, C., Tsung-Yi, H.: Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24658–24667 (2023)

[58] Nicolas, P., Patrick, M., Xi, W., Somesh, J., Ananthram, S.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 582–597 (2016). IEEE

[59] Xiaoyu, C., Zhenqiang, G.N.: Mitigating evasion attacks to deep neural networks via region-based classification. In: Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 278–287 (2017)

[60] Hyeungill, L., Sungyeob, H., Jungwoo, L.: Generative adversarial trainer: Defense to adversarial perturbations with gan. arXiv preprint arXiv:1705.03387 (2017)

[61] Yunseok, J., Tianchen, Z., Seunghoon, H., Honglak, L.: Adversarial defense via learning to generate diverse attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2740–2749 (2019)

[62] Mazda, M., Soheil, F.: Sample efficient detection and classification of adversarial attacks via self-supervised embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7677–7686 (2021)

[63] Zhezhi, H., Siraj, R.A., Deliang, F.: Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 588–597 (2019)

[64] Hao-Yun, C., Jhao-Hong, L., Shih-Chieh, C., Jia-Yu, P., Yu-Ting, C., Wei, W., Da-Cheng, J.: Improving adversarial robustness via guided complement entropy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4881–4889 (2019)

[65] Aamir, M., Salman, K., Munawar, H., Roland, G., Jianbing, S., Ling, S.: Adversarial defense by restricting the hidden space of deep neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3385–3394 (2019)

[66] Yaoyao, Z., Weihong, D.: Adversarial learning with margin-based triplet embedding regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6549–6558 (2019)

[67] George, C., Calvin, M., Simon, L.: Architectural adversarial robustness: The case for deep pursuit. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7150–7158 (2021)

[68] Gaojie, J., Xinping, Y., Dengyu, W., Ronghui, M., Xiaowei, H.: Randomized adversarial training via taylor expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16447–16457 (2023)

[69] Nilaksh, D., Madhuri, S., Shang-Tse, C., Fred, H., Li, C., E, K.M., Horng, C.D.: Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. arXiv preprint arXiv:1705.02900 (2017)

[70] Chuan, G., Mayank, R., Moustapha, C., Laurens, V.D.M.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)

[71] Meng, D., Chen, H.: Magnet: a two-pronged defense against adversarial examples. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147 (2017)

[72] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1787 (2018)

[73] Song, Y., Kim, T., Nowozin, S., Ermon, S., Kushman, N.: Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. arXiv preprint arXiv:1710.10766 (2017)

[74] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer

[75] Bai, Y., Feng, Y., Wang, Y., Dai, T., Xia, S.-T., Jiang, Y.: Hilbert-based generative defense for adversarial examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4784–4793 (2019)

[76] Moosavi-Dezfooli, S.-M., Shrivastava, A., Tuzel, O.: Divide, denoise, and defend against adversarial attacks. arXiv preprint arXiv:1802.06806 (2018)

[77] Sun, B., Tsai, N.-h., Liu, F., Yu, R., Su, H.: Adversarial defense by stratified convolutional sparse coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11447–11456 (2019)

[78] Gupta, P., Rahtu, E.: Ciidefence: Defeating adversarial attacks by fusing class-specific image inpainting and image denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6708–6717 (2019)

[79] Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 501–509 (2019)

[80] Zhou, D., Wang, N., Peng, C., Gao, X., Wang, X., Yu, J., Liu, T.: Removing adversarial noise in class activation feature space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7878–7887 (2021)

[81] Nunes, D.C.L., Jonathan, T.: Artificial immune systems: a new computational intelligence approach. Springer Science & Business Media (2002)

[82] Chandrasekaran, M., Asokan, P., Kumanan, S., Balamurugan, T., Nickolas, S.: Solving job shop scheduling problems using artificial immune system. The International Journal of Advanced Manufacturing Technology **31**(5-6), 580–593 (2006)

[83] Cutello, V., Nicosia, G., Pavone, M., Timmis, J.: An immune algorithm for protein structure prediction on lattice models. IEEE Transactions on Evolutionary Computation **11**(1), 101–117 (2007)

[84] Peilan, L.T.X.: A clonal selection algorithm for dynamic multimodal function optimization. Swarm and Evolutionary Computation **50** (2019)

[85] I., H.-J.L., Hao, X.-H., Zhang, L.: Clonal selection algorithm for multi-objective optimization. ence Technology & Engineering, 453–482 (2008)

[86] Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586–587 (1991). IEEE Computer Society

[87] Kaiming, H., Haoqi, F., Yuxin, W., Saining, X., Ross, G.: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

[88] Ziwei, L., Ping, L., Xiaogang, W., Xiaoou, T.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (2015)

[89] Huang, G.B., Mattar, M., Berg, T., Eric, L.-M.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008)

[90] Ira, K.-S., M., S.S., Daniel, M., Evan, B.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4873–4882 (2016)

[91] Transferable Adversarial LFW. http://www.whdeng.cn/TALFW/index.html

[92] Komkov, S., Petiushko, A.: Advhat: Real-world adversarial attack on arcface face id system. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 819–826 (2021). IEEE

[93] Zhu, Z.-A., Lu, Y.-Z., Chiang, C.-K.: Generating adversarial examples by makeup attacks on face recognition. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 2516–2520 (2019)