FAIC-Attack: An Adversarial Watermarking Attack against Face Age based on Identity Constraint

Xuning Zheng
zhengxuning@seu.edu.cn
School of automation, Southeast University
Nanjing, China

Xiankang Wang*
213210973@seu.edu.cn
School of Economics and Management, Southeast University
Nanjing, China

Ziyi Yu
213213212@seu.edu.cn
School of automation, Southeast University
Nanjing, China

Siyu Xia
xsy@seu.edu.cn
School of automation, Southeast University
Nanjing, China

Abstract

Recently, there has been increasing concern about the security of facial recognition systems, especially in the context of black-box attacks. As attackers continue to devise new ways to exploit vulnerabilities, attention to age estimation in facial recognition becomes critical. Age estimation is also a critical task for a variety of applications, evolving with advances in computer vision and deep learning. In this paper, an identity-constrained face age against Watermark attack (FAIC) method based on DDE algorithm is proposed. The method finds the optimal solution of watermark addition by changing the position of the watermark in the host image, the transparency of the watermark, the size of the watermark, and the rotation angle to deceive deep neural networks. In addition, we also try to constrain the face identity in the attack to achieve the effect of only changing the face age without changing the face identity after the attack. A series of experiments show that our method can improve the stability of face identity while attacking face age, and improve the success rate of attack by changing the watermark size and rotation angle, which proves that our added parameter settings are effective. The proposed FAIC method and the constraint on face identity provide an effective and stable method for the black box attack of face age estimation.

*Co-first author

ICMVA 2024, March 10–12, 2024, Singapore, Singapore

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts: • **Computing methodologies** → **Computer vision problems**.

Keywords: Adversarial Examples, Watermark Perturbation, Double Differential Evolution, Age Estimation

ACM Reference Format:
Xuning Zheng, Xiankang Wang, Ziyi Yu, and Siyu Xia. 2023. FAIC-Attack: An Adversarial Watermarking Attack against Face Age based on Identity Constraint. In *Proceedings of (ICMVA 2024).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

With the continuous development of artificial intelligence (AI) technology, facial recognition systems [1] have found widespread applications in various aspects of social life, including but not limited to security checks, financial transactions, and intelligent access control. However, as facial recognition technology advances, attackers are constantly searching for new methods to deceive these systems, leading to potential security breaches [2]. The widespread use of facial recognition systems has raised concerns about their security and robustness. In this context, black-box attacks [3] have become a focal point of research for scholars and security experts. Black-box attacks involve attackers having limited information access to the target system, without knowledge of the specific structure and parameters of the system. In black-box attacks, attackers need to design sophisticated methods to exploit vulnerabilities in facial recognition systems through limited trial and error, deceiving the system. To enhance the success rate and deception of black-box attacks, researchers have been exploring various methods and technologies.

In the realm of facial recognition technology, age estimation [4] is a crucial task with significant applications in many scenarios. Age estimation involves the analysis and identification of facial features through advanced technologies such as computer vision, deep learning, and artificial intelligence to infer the age range of an individual. As technology

advances, age estimation techniques have evolved from traditional image processing methods to more accurate and precise methods, especially with the application of convolutional neural networks. Modern age estimation algorithms can analyze facial details more finely, considering multiple feature points and contextual information for a more accurate estimation of an individual's age. This technology finds widespread applications in social networks, security monitoring, and medical imaging, providing convenient and efficient means of identity verification and individual recognition. However, technological advances have also sparked discussions about privacy and ethical concerns, necessitating regulatory oversight and guidance on relevant laws and ethical standards alongside technological applications.

Adversarial attack [5], [6] refers to techniques aimed at machine learning models, using intentionally designed input data to mislead the model's output or cause a decline in its performance. The goal of such an attack is to manipulate the input data subtly, resulting in erroneous classification or predictions by the model, revealing weaknesses and disrupting its normal functioning. Adversarial attack can be implemented by adding small, carefully designed perturbations that are nearly imperceptible to the human eye but sufficient to deceive machine learning models. This type of attack requires an in-depth understanding of the model to identify its sensitivity to input and bypass its defense mechanisms through clever adjustments to input data. Research on adversarial attacks is crucial to enhance the robustness and security of machine learning models to address potential threats in practical applications.

Watermarking is a technique that involves the embedding of information, usually in a transparent or semi-transparent form, into digital images, documents, or videos to identify ownership or origin. Watermarks, in the form of text, icons, or graphics, exist inconspicuously within the content but can be identified and verified through specific methods when needed. Watermarks serve as a means of protecting intellectual property and preventing unauthorized reproduction. They offer a simple yet effective way for creators to ensure the intellectual property of their works without disrupting the viewing experience. Additionally, watermarks can be used to track instances of piracy, providing legal grounds for copyright protection. Furthermore, watermarks serve as an anti-counterfeiting measure, ensuring the authenticity and credibility of images, documents, or videos [7], [8]. Thus, watermark technology plays a crucial role in digital content protection and copyright management. Here, we chose to use the BHE algorithm [9] to add watermarks with specific transparency to particular positions in the image. The BHE algorithm is designed to determine the appropriate transparency of the watermark image and embed it into the host image's specific location. The BHE algorithm employs multiple initial points and crossover operations, facilitating the identification of the global optimum and increasing the success rate of the attack.

In this paper, we explore the relationship between the size of the watermark and the success rate of the attack in the adversarial attack. Furthermore, we introduce two new parameters, β and γ (where β is the watermark size scaling factor and γ is the watermark rotation angle), to optimize attack on facial age attributes. In our experiments, we use five key parameters to adjust the size and orientation of the watermark and attack the age estimation model with the adjusted watermark. We employ the basin hopping algorithm [9] to generate adversarial attack samples, aiming to find the optimal solution for the attack. Our method is a black-box attack, requiring only access to the classifier's output category and confidence to achieve the attack. Using this approach, we can generate effective attack samples that easily evade existing adversarial attack evaluation standards [10].

Simultaneously, we also investigate whether our attack leads to changes in facial identity labels and conclude, through statistical analysis, that the attack may cause changes in facial identity labels. However, in adversarial attack, our goal is to attack facial age while maintaining the stability of facial identity labels. To achieve this objective, we propose an identity-constrained face age against Watermarking attack (FAIC) based on the DDE algorithm. We provide a detailed explanation of the design principles and workings of the constraint condition, in order to minimize interference with facial identity information while maximizing the effectiveness of adversarial attack. By limiting the scope of the impact of the attack, our method achieves adversarial attack while maintaining the stability of facial identity labels. In summary, by introducing the constraint condition that restrict the impact on facial identity labels during the attack, we enhance the effectiveness of adversarial attack on facial age without changing facial identity labels.

In general, the main contributions of this work include:

1. We propose an attack method targeting facial age called FAIC-Attack. By embedding specific transparent watermarks in the attack and finding the global optimal solution, we successfully attack the facial age estimation model and provide a new application scenario for black box attacks.

2.We have introduced two new parameters, watermark scaling factor (β) and watermark rotation angle (γ). Adjusting these parameters to optimize the attack effect on facial age, improve the success rate and deception of attacks, and optimize the robustness of adversarial attacks is an innovative point in the field of adversarial attacks.

3. We use DDE to generate counterattack samples, aiming to find the optimal solution of the attack. Through statistical analysis, pay attention to whether the attacks cause face identity tags to change, and propose an identity-constrained face age against Watermarking attack (FAIC) based on the DDE algorithm to ensure that face identity does not change as much as possible while confronting the attacks. It is the

first known adversarial watermarking method to successfully attacks face age without changing face identity.

2 Related Works

2.1 Adversarial Attacks

The objective of adversarial attacks is to deceive a deep learning model by using adversarial samples. Depending on the desired outcome, adversarial attacks can be classified into targeted attacks and untargeted attacks. Targeted attacks aim to classify the adversarial sample into a specific category, whereas untargeted attacks seek to disrupt the model's output, making it inconsistent with the output of the clean image. Depending on whether the attacker can access information about the model, adversarial attacks can be categorized as white-box attacks and black-box attacks. In a white-box attack, when the attacker possesses some or all of the model's information, such as its structure, they have access to this information to execute the attack. Goodfellow et al. [6] introduced the FGSM method for generating adversarial samples in white-box scenarios, which is based on the changes in the gradient direction of deep neural networks. The I-FGSM [11] method has made certain improvements. Conversely, in a black-box attack, the attacker lacks access to the model's internal information and can only obtain model outputs. Specific black-box attack methods will be introduced in the following.

2.2 Black-Box Attacks

Typically there are three approaches for black-box attacks: transfer attacks, where adversarial samples generated in a white-box context are transferred to a black-box scenario; query-based attacks, and score-based attacks. MI-FGSM [12] , DI-FGSM [13], TI-FGSM [14], PI-FGSM [15] are all transfer attacks based on the FGSM method. In [16], Jiawei Su et al. proposed a single-pixel black-box score-based attack method using differential evolution, with the only available information being the probability label. Sparse-RS [17]is a score-based sparse black-box attack method, which doesn't restrict the magnitude of modifications but only alters a small portion of pixels. Optimization Attack [18] is the first decision-based attack that guarantees convergence. This method formulates the black-box attack problem as a real-valued optimization problem.

2.3 Watermarking Attacks

Xiaojun Jia et al. introduced the adv-watermark [9] method, which generates meaningful adversarial samples by adding watermarks to clean data. This approach utilizes population-based optimization algorithms to find the optimal adversarial samples. FAWA [10] is an adversarial watermark sample generation method based on the differential evolution algorithm. This method successfully attacked the Aliyun Image Recognition API, although it required a relatively long time to execute the attack. In [19], Xianyu Zuo et al. introduced an adversarial watermark sample generation method based on the Particle Swarm Optimization algorithm, known as MISPSO-Attack, which outperforms its predecessor in terms of performance.

2.4 Face Recognition and Face Age estimation

DeepFace [20] and DeepID [21] treat face recognition as a multi-class classification problem. They employ deep CNNs to learn features supervised by the softmax loss. In order to increase the Euclidean margin between classes in the feature space, Triplet loss [22] and center loss [23] are proposed. SphereFace [24] introduces the angular softmax loss to learn angularly discriminative features. CosFace [25] maximizes the cosine margin by utilizing the large margin cosine loss. ArcFace [26] proposes the additive angular margin loss for learning highly discriminative features. Age estimation based on facial images is a "special" pattern recognition problem: on the one hand, since each age value can be regarded as a class, age estimation can be viewed as a classification problem [27] utilizes an age segmentation method to address the issue of inaccurate classification due to the continuity of age. Due to the continuity of age, age estimation can also be treated as a regression problem [28]. The DLDL[29]method converts real age values into a discrete age distribution to adapt to the entire age distribution.

3 FAIC-Attack: An Adversarial Watermarking Attack against Face Age based on Identity Constraint

In this section, we introduce the proposed method from three aspects: problem formulation,perceivable color watermark generation and double differential evolution constrained by face identification. The framework of our work is shown in Figure1.

3.1 Problem Formulation

The generate adversarial perturbations with perceptible watermarks is a method to modified images, which perturbs the age prediction of CNN but retains the identical information of human faces. In our algorithm, (x, y) position, transparency, scaling and rotation angle affect the result of the age and identification. Therefore, in order to achieve our goal, the process of algorithm can be seen as an optimization problem with a constraint. According to the subsection *Perceivable Color Watermark Generation*, we could define the process of attack as $\mathcal{A}(I, W, x, y, \alpha, \beta, \gamma)$, which generate a new image $I_{new}(M, N)$ based on the target image $I_{origin}(M, N)$. The new image can be the input of CNN classifier with two outputs of probable vectors:

$$P_{age} = [p_0, p_1, \cdots, p_{100}] \tag{1}$$

$$P_{id} = [p_{name,1}, p_{name,2}, \cdots, p_{name,n}] \tag{2}$$

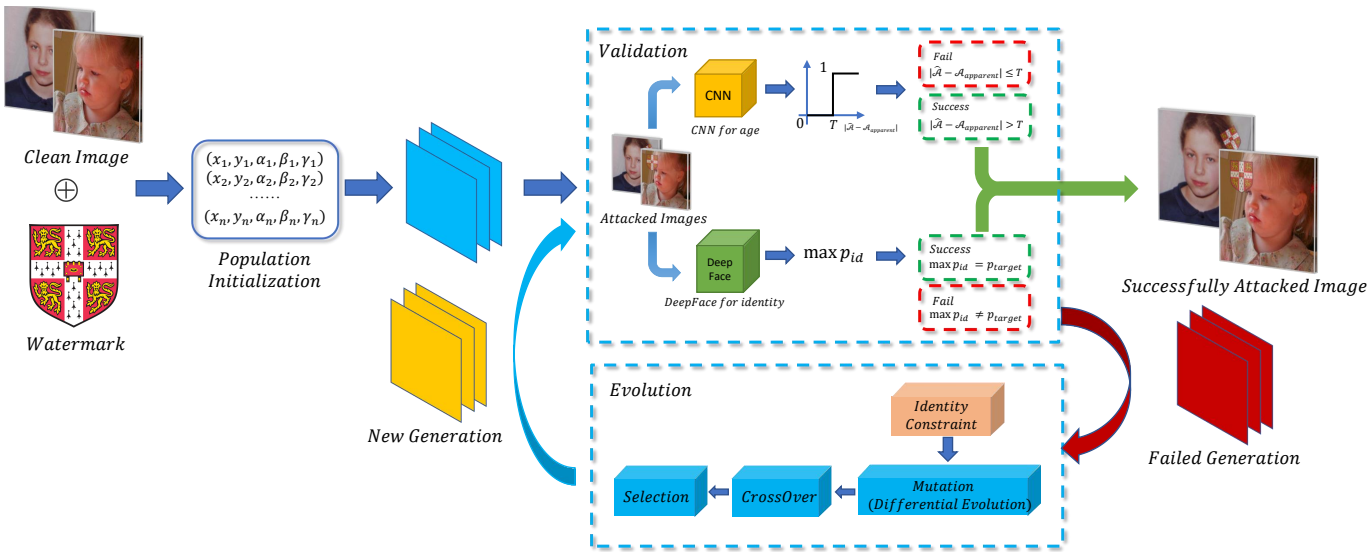

Figure 1. The figure shows the process of FAIC-attack to generate the adversarial samples. The initial solutions generated by Guass Distribution. This set of pending solutions is optimized based on the well-trained face age classifer's output under face identity constrains.

where P_{age} is the output vector of the age estimate classifier and P_{id} is the output of the face identification classifier. The aim of our attack is to minimize the age of our target image p_{target}, but control the identification of target unchangeable, which can be describe as formulations below:

$$\begin{cases} \hat{\mathcal{A}} = \arg\min_{\mathcal{A}} P_{age} \\ \max p_{id} = p_{target} \end{cases} \tag{3}$$

Variables of this problem are as follow: 1) The center of the watermark (x, y) in the original image; 2) The transparency of the watermark α; 3) The scaling factor β, which can control the size of watermark; 4) The rotation angle γ, which shows the orientation of the watermark. Embedding the adversarial watermark into original image can protect the safety of targeted personal information but doesn't affect the aesthetics of the image. At the same time, watermark has practical significance such as contractual protection, etc. Given in the real physical world, a small gap of ages may not change the appearance of a person, we give a hyperparameter T, as the threshold to estimate the result of our attack. Therefore, we can define the attack success or not as:

$$R = \begin{cases} Success, & \text{if } |\hat{\mathcal{A}} - \mathcal{A}_{apparent}| > T \text{ and } \max p_{id} = p_{target}; \\ Fail, & \text{else} \end{cases} \tag{4}$$

3.2 Perceivable Color Watermark Generation

We use alpha blending [30] to add watermark to original image. Alpha channel refers to the transparency of the foreground region compared to the background. In this paper, we use α to represent the value of the alpha channel, $I(M, N)(I)$ is the original image, M and N to represent the height and weight of the image, $W(m, n)$ to represent the watermark whose size is m, n, $I^{new}(M, N)(I^{new})$ to represent the generated image with a watermark. $v(x)$ is an image x and i, j is the pixel position, where $i \in (x, x+m), y \in (y, y+n), x, y$ to represents the position where the watermark is embedded. Meanwhile, we also consider the scale β and angle γ of the watermark. $r(x, \gamma)$ denotes the rotation of an image x, So the generation for $I^{new}(M, N)$ is formulated as:

$$v\left(I^{new}\right) = \frac{v\left(r(W, \gamma)\right)_{i-x, j-y} * \alpha + v\left(I\right)_{i,j} * (255 - \alpha)}{255} \tag{5}$$

when $i \notin (x, x+m), y \notin (y, y+n)$, $I^{new}(M, N)$ is formulated as :

$$v\left(I^{new}\left(M, N\right)\right)_{i,j} = v\left(I\left(M, N\right)\right)_{i,j} \tag{6}$$

And we also want to change the scale of the watermark,the process can be formulated as:

$$\begin{cases} \eta = \min\left(\frac{M*\beta}{m}, \frac{N*\beta}{n}\right), \\ m_\beta = m * \beta, n_\beta = n * \beta. \end{cases} \tag{7}$$

In this paper, we use University of Cambridge logo as image watermarks.

3.3 Double Differential Evolution constrained by face identification

Based on Basin Hopping Evolution(BHE)[9], we propose a global optimization algorithm, which is called Double Differential Evolution(DDE). Our method is a group evolution algorithm based on Differential Evolution, which can be used for finding the global minimum of a multivariate function. As shown in Figure 1, DDE includes population initialization,

Differential Evolution(DE), crossover and selection operations. During each iteration, we use DE to produce a set of better solutions(We call this procedure Mutation) and conduct crossover operations to generate a new set of candidate solutions. And than in selection operation, if the candidate solutions posses the smaller multivariate function value, they survive and are passed to the next generation.

3.3.1 Population Initialization

Firstly, We would like to recast the optimization problem we want to solve:

$$\begin{cases} \underset{e(x,y,\alpha,\beta,\gamma)}{arg\min} \ f\left(g\left(H,W,x,y,\alpha,\beta,\gamma\right)\right) \\ s.t.\alpha \leqslant L \end{cases} \tag{8}$$

where $g(H, W, x, y, \alpha, \beta, \gamma)$ is the visible watermark algorithm,and $f(\cdot)$ is the well-trained age classify. DDE algorithm is an optimization algorithm based on group evolution. We regard each solutions as an individual of a population. And the elements($x, y, \alpha, \beta, and\ \gamma$) are considered as its genes. Let $X_{i,g}$ denote the i-th individual in the g-th generation population. And $X_{i,g,j}(j = 0, 1, 2, 3, 4)$ denotes the j-th gene of $X_{i,g}$. Therefore, we initialize a popualtion as follows:

$$X_{i,0,j} = X_{\min,j} + \mathcal{N}(0,1) \cdot \left(X_{\max,j} - X_{\min,j}\right), j = 0, 1, 2, 3, 4 \tag{9}$$

where $X_{i,0,j}$ is the j-th gene of the i-th individual in the initial population, $X_{min,j}$ is the minimum of the j-th gene and $X_{max,j}$ is the maximum of the j-th gene.

3.3.2 Differential Evolution

Our algorithm's framework is Differential Evolution Algorithm. To find more efficient solutions, we use another Differential Evolution Algorithm to act as mutation operation. The details of Differential Evolution Algorithm can be found in [31]

3.3.3 Crossover

As for the current solution $X_{i,g}$ and the corresponding DE optimization solution $V_{i,g}$, we conduct crossover operation to get a candidate solution $U_{i,g}$. It is formulated as

$$U_{i,g,j} = \begin{cases} V_{i,g,j}, \ \mathcal{N}(0,1) \leqslant CR, \\ X_{i,g,j}, \ else \end{cases} \tag{10}$$

where $U_{i,g,j}$ is the j-th gene of $U_{i,g}$,$V_{i,g,j}$ is the j-th gene of $V_{i,g}$,$X_{i,g,j}$ is the j-th gene of $X_{i,g}$ and CR is the crossover probability which represents the degree of information exchange in the population evolution. It is a super parameter.

3.3.4 Selection

We adopt a greedy selection strategy to select a better solution as the next generation solution. We can formulated the operation as:

$$X_{i,g+1} = \begin{cases} U_{i,g}, \ f\left(U_{i,g}\right) \leqslant f\left(X_{i,g}\right) \\ X_{i,g}, \ else \end{cases} \tag{11}$$

What's more, due to the transferability of the adversarial samples[32], there's a high probability that the samples we generate will change the identity of the face. In order to make the attack only target the age attribute of the face, we add a constraint on the identity of the face. The details of DDE algorithm is given in Algorithm 1.

Algorithm 1: Double Differential Evolution constrained by face ID

Input: Population :P; Dimension: 5; Generation: G; Iteration: I; A well-trained face-ID classify F

Output: The best solution - Δ

```
g ← 0;
for i = 1 to P do
    for j = 1 to 5 do
        X_{i,0,j} = X_{min,j} + rand(0,1) · (X_{max,j} − X_{min,j})
    end
end
while f_t(Δ) ≥ ε and g ≤ G do
    for i = 1 to P do
        V_{i,g} = DifferentialEvolution(X_{i,g}, I)
        for j = 1 to 5 do
            U_{i,g,j} = Crossover(V_{i,g,j}, X_{i,g,j})
        end
    end
    if f_t(U_{i,g}) ≤ f_t(X_{i,g}) then
        X_{i,g} = U_{i,g}
        if f_t(U_{i,g}) ≤ f_t(X_Δ) then
            Δ = X_{i,g}
        end
    else
        X_{i,g} = X_{i,g}
    end
    if F(I_Δ) ≠ F(I) then
        break;
    end
    g ← g + 1
end
```

4 Experiment

4.1 Experiment Environment

All the experiments are carried out using 24 vCPU AMD EPYC 7642 48-Core Processor and a Nvidia RTX 3090 GPU. We use APPA-REAL (real and apparent age) as the dataset for our experiments[33], which contains 7,591 images with associated real and apparent age labels. The total number of apparent votes is around 250,000. On average we have around 38 votes per each image and this makes the average apparent age very stable (0.3 standard error of the mean). And the images are split into 4113 train, 1500 valid and 1978 test images. We tested several publicly available pre-trained models, including Resnext-32x4d[34], AlexNet[35], GoogLeNet[36], VGG-16[37] and ShuffleNet[38]. We did not retrain and fine-tune these networks

4.2 Experiments of attacks without the identical constraint

4.2.1 Hyper-parameters Selection

We conduct a large number of experiments to determine two hyper parameters in DDE. One is the number of mutation iterations I, the other one is crossover probability CR. We adopt DDE to attack DNN models using University of Cambridge logo. In detail, we compute the attack success rates of the Resnext-32x4d on 400 random image of the APPA-REAL dataset. The result is shown in Table 1.From Table 1, it is clear that the attack success rate increases when I increases. That is, as the number of mutation iterations increases, the solution generated by DE will be better, resulting in achieving a higher attack success rate. But more iterations mean more time spent. Considering time complexity, we set CR to 0.9 and I to 3. In this way, DDE can achieve the highest attack success rate 61.5%.

Table 1. Selection of hyper-parameters in DDE

	I = 2	I = 3	I = 4	Average
CR = 0.8	58%	64.80%	62.70%	61.83%
CR = 0.9	60.70%	61.50%	64%	62.07%
CR = 1.0	62.80%	64.70%	68%	65.17%
Average	60.50%	63.67%	64.90%	65.03%

4.2.2 Attack Performance

The success rate in the unconstrained case is given in the Table2.For Resnext-32x4d network, the impact of the samples generated by the attack on the accuracy of different networks is illustrated in Table 2. It can be observed that the attack generated based on Resnext-32x4d has a strong attacking effect on itself. This attack has transfer capabilities. For example, it can also fool AlexNet network with a success rate of 43.5%.

Table 2. The success rate in the unconstrained case

Model	Resnext Attack	ShuffleNet Attack
Resnext_32x4d	60.79%	34.23%
AlexNet	43.15%	41.85%
GoogLenet	38.01%	39.6%
VGG-16	45.33%	42.35%
ShuffleNet	40.98%	65.47%

4.2.3 Attack Performance on CAM

After getting the results, we use CAM(Class Activation Mapping) of Resnext to make a further explanation on our work, shown in Figure 2. CAM can show the regions of interest in the model. In the figure, we give three examples. The first column is original images, second column is CAMs of them. Then the third column is attacked images, and CAMs of them at last. As we can see, after attacked, the regions of interest in every images moved, which led to wrong prediction of age. This is consistent with our desire about attacking target person's age.

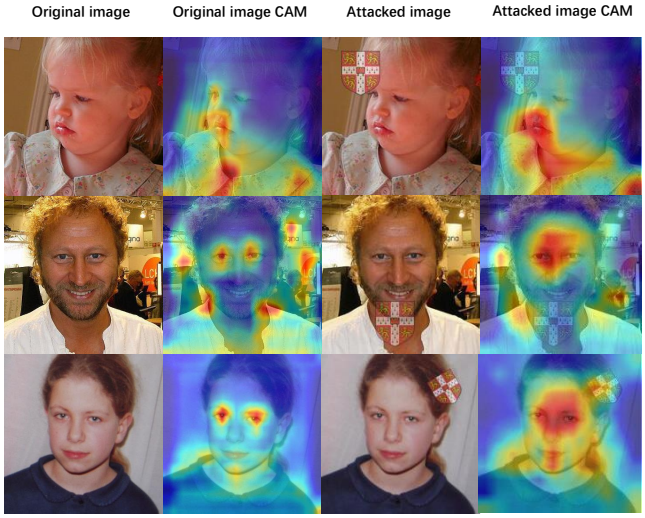

Figure 2. We use CAM of Resnext to show the regions of interest of the original image and the attacked image. We can find that after attacked, the regions of interest in every images moved, which led to wrong prediction of age.

4.3 Experiments of effect on face identification based on age attack

Before we start our validation of constrained attack's performance, we need to test how our embedded watermarks affect the identification of images. Under APPA-REAL dataset, we did an experiment of effect on face identification based on age attack. We input the image we attack to DeepFace, and use several models to check whether new images can be judge as same person by different models. As we can see in the Figure 3, when we set our target is only age attack, our success rate is around 60%. However, if the target is set for age attack and identity protection, the success rate is around 40%, which shows watermarks can also affect the prediction of target identity. Under cosine metric, we can get the distance between new and origin images, such as in Figure 4.

We also use different models to check the success rate when we need to attack age but to protect personal identity. It can be seen from the data, if we don't use some methods to constrain the effect of watermark on targets image. There are about 33% percentage that the identity will be changed, which gets more personal information lost and can not protect or attack some of personal information precisely.

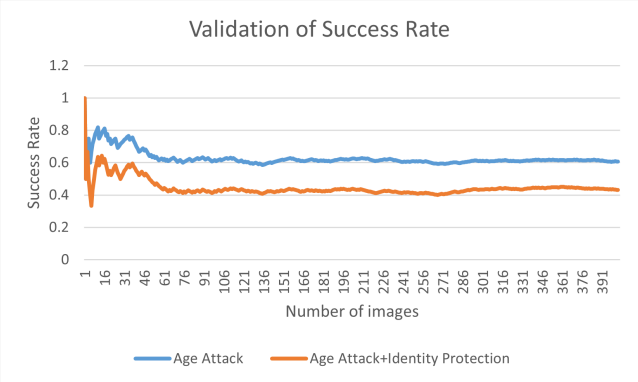

Figure 3. Success rate at the target of age attack or age attack+identity protection(VGG-Face), when we set our target is only age attack, our success rate is around 60%. If the target is set for age attack and identity protection, the success rate is around 40%

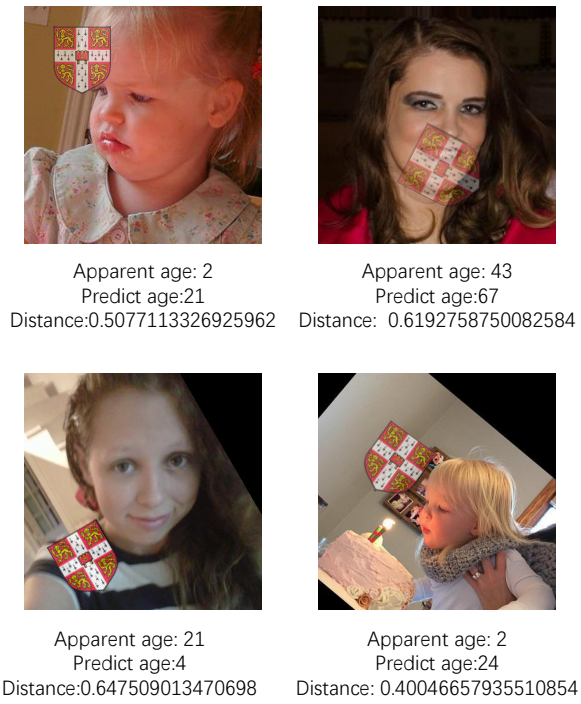

Figure 4. Under cosine metric, we can get the distance between new and origin images. When the ditance is bigger than 0.4, The identity of the face changes

4.4 Attack Performance under the identity constraint

In this section, we use the same models and attack method as Section 4.2 to evaluate the performance with and without the identity constraint. The attacks under constrained and unconstrained identities are shown in the Table 4 & 5.

Table 3. Success rate on different models

Model	LFW Score	Success Rate
Facenet512	99.65%	44.13%
Arcface	99.41%	45.59%
Dlib	99.38%	43.98%
Facenet	99.20%	39.72%
VGG-Face	98.78%	43.17%
Human-beings	97.53%	40.65%
OpenFace	93.80%	35.10%

Table 4. Unconstrained success rate (age+id) The DeepFace Arcface model was used to align identities

Model	Resnext Attack	ShuffleNet Attack
Resnext_32x4d	43.17%	17.08%
AlexNet	20.93%	24.87%
GoogLenet	27.75%	19.56%
VGG-16	25.32%	25.86%
ShuffleNet	21.64%	38.09%

Table 5. Success rate under the constraint (age+id) The DeepFace Arcface model was used to align identities

Model	Resnext Attack	ShuffleNet Attack
Resnext_32x4d	50.8%	20.77%
AlexNet	25.23%	26.42%
GoogLenet	30.91%	23.24%
VGG-16	33.78%	31.16%
ShuffleNet	29.92%	50.93%

From Table 4 and Table 5, we can find the success rate under the constraint is commonly higher than unconstrained one. The result shows that, our attack method is specific, which means attack can change the age of target person, but keep its identity unchanged. At the same time, our attack method is also robust, the adversarial samples from Resnext Attack and ShuffleNet Attack can be effective in other models.

5 Conclusion

In this paper, we propose a novel attacking method called FAIC to fool well-trained DNN model. The method finds the optimal solution by changing the position of the watermark in the clean image, the size of the watermark and the rotation angle of the watermark. In addition, we also try to constrain the face identity in the attack to achieve the effect of only changing the age of the clean image without changing the face identity, which we called specific attack. And the proposed method could be more commonly used in the real world.

6 Discussion

The method we proposed can be used in Face privacy protection.One of the application scenarios we envision is In some places, there is an age limit, such as greater or less than 18 years old, and the data stored in our database is watermarked to protect age privacy. At the same time, our method can also be extended to multiple face attributes such as gender and expression.

References

[1] Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. *journal of information processing systems*, 5(2):41–68, 2009.

[2] Yi Zeng, Han Qiu, Gerard Memmi, and Meikang Qiu. A data augmentation-based defense method against adversarial attacks in neural networks. In *Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part II 20*, pages 274–289. Springer, 2020.

[3] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[4] Prachi Punyani, Rashmi Gupta, and Ashwani Kumar. Neural networks for facial age estimation: a survey on recent advances. *Artificial Intelligence Review*, 53:3299–3347, 2020.

[5] Han Qiu, Tian Dong, Tianwei Zhang, Jialiang Lu, Gerard Memmi, and Meikang Qiu. Adversarial attacks against network intrusion detection in iot systems. *IEEE Internet of Things Journal*, 8(13):10327–10335, 2020.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Christophe De Vleeschouwer, J-F Delaigle, and Benoit Macq. Invisibility and application functionalities in perceptual watermarking an overview. *Proceedings of the IEEE*, 90(1):64–77, 2002.

[8] Gordon W Braudaway. Protecting publicly-available images with an invisible image watermark. In *Proceedings of international conference on image processing*, volume 1, pages 524–527. IEEE, 1997.

[9] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Xiaoguang Han. Adv-watermark: A novel watermark perturbation for adversarial examples. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1579–1587, 2020.

[10] Hao Jiang, Jintao Yang, Guang Hua, Lixia Li, Ying Wang, Shenghui Tu, and Song Xia. Fawa: Fast adversarial watermark attack. *IEEE Transactions on Computers*, 2021.

[11] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[13] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[14] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[15] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 307–322. Springer, 2020.

[16] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[17] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6437–6445, 2022.

[18] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.

[19] Xianyu Zuo, Xiangyu Wang, Wenbo Zhang, and Yadi Wang. Mispso-attack: An efficient adversarial watermarking attack based on multiple initial solution particle swarm optimization. *Applied Soft Computing*, 147:110777, 2023.

[20] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[21] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[23] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016.

[24] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[25] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[26] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[27] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015.

[28] Yunxuan Zhang, Li Liu, Cheng Li, et al. Quantifying facial age by posterior of age comparisons. *arXiv preprint arXiv:1708.09687*, 2017.

[29] Bin-Bin Gao, Xin-Xin Liu, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Learning expectation of label distribution for facial age and attractiveness estimation. *arXiv preprint arXiv:2007.01771*, 2020.

[30] Bo Shen, Ishwar K Sethi, and Vasudev Bhaskaran. Dct domain alpha blending. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, volume 1, pages 857–861. IEEE, 1998.

[31] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11:341–359, 1997.

[32] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[33] R Timofte E Agustsson, X Baro S Escalera, and R Rothe. I Guyon. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*. IEEE, 2017.

[34] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.