TARGETED ANONYMIZATION: A FACE IMAGE ANONYMIZATION METHOD FOR UNAUTHORIZED MODELS

Kangyi Ding, Teng Hu, Xiaolei Liu, Weina Niu, Yanping Wang and Xiaosong Zhang✉

School of Computer Science and Engineering
Institute for Cyber Security
University of Electronic Science and Technology of China, Chengdu, China

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859898

ABSTRACT

As an important biometric feature of every person, face data has faced serious risks of leakage in recent years. Lawbreakers can use face recognition systems (FRS) to analyze the leaked face data and then correlate other private information, causing serious privacy leaks. For security reasons, we hope our face images can only be recognized by the organizations' authorized models. To achieve this goal, this work proposes a targeted face image anonymization method that only enables anonymization for unauthorized facial recognition models, whilst authorized models, human eyes can still accurately recognize faces. Our method mainly uses transfer-based adversarial attacks to achieve anonymization. On this basis, we propose constraints for generating targeted anonymization samples and boundary walking strategy, focusing on improving the anonymization for unauthorized models while guaranteeing the accurate recognition of authorized models. Local experiments prove that our method can reduce the recognition probability of unauthorized models while guaranteeing the correctness of authorized models. Finally, we apply our approach to an online face recognition API and experimentally demonstrate that our approach can significantly reduce the recognition accuracy of the commercial face recognition model.

Index Terms**—** Face anonymization, targeted anonymization, adversary, transferability

1. INTRODUCTION

Today, in the context of the popularity of social networking platforms, unauthorized trading of personal information is common on the Internet. The face data kept and managed by governments and enterprises often have higher authenticity and thus higher acquisition value. Once a face data breach has occurred, by using artificial face recognition algorithms, offenders can easily achieve accurate matching with users' data on the Internet and social networking platforms. So they can establish more perfect social networks and users profile, resulting in serious leakage of users' private information. All of this undoubtedly poses challenges for governments and enterprises face data storage.

Governments and enterprises need face data to perform face recognition. For example, governments store photos of faces recorded by surveillance devices in specific scenes, and the photo on the ID card is often used as face recognition for security check scenarios, but also needs to retain the functionality of being recognized by the human eyes for other scenarios. Enterprises want to publish images that can only be recognized by their face recognition algorithms for commercial copyright reasons. Therefore, if organizations keep face data that can only be recognized by their models (authorized models), it will significantly reduce the harm caused by data breaches. We propose a targeted face image anonymization method. The face generated by our method can be anonymized to the FRS of unauthorized organizations, while can be recognized and matched correctly by the FRS of authorized parties. The method adds the smallest possible perturbations to face images to retain more information about the face.

Current research on face image anonymization for FRS is relatively limited, and the common strategy is to change the original image. Typical methods used for traditional face image anonymization include image processing [1], K-anonymization algorithms [2], generative adversarial networks [3, 4], and adversarial attacks. Specifically, adversarial attacks can achieve anonymization to FRS by adding very small perturbations and exploiting the transferability of the adversarial samples with little change in the semantic information of the images [5, 6]. However, none of the above approaches can achieve differentiated anonymity between authorized and unauthorized models.

For the need to achieve anonymization to unauthorized models and non-anonymization to authorized models, we propose generating constraints and boundary walking. These strategies preserve the adversarial transferability of the generated samples as much as possible while guaranteeing the correct recognition rate of the target model.

In summary, we make the following contributions: (1) To the best of our knowledge, we are the first to propose an anonymization method that is valid only for the unauthorized FRS. (2) We propose constraints to ensure the generated samples can be used as gallery samples to match with other samples by the authorized model. (3) We propose a boundary walking strategy to enhance the success rate of adversarial attacks against unauthorized models while guaranteeing the correct recognition rate of authorized models. Experimentally, we demonstrate that our method can preserve more adversarial transferability to the unauthorized models, but has almost no impact on the recognition accuracy of the authorization model. (4) We apply our method to online face recognition API attacks. The results show that our method can significantly reduce the correct recognition rate of API.

2. RELATED WORK

In terms of adversarial attacks against face recognition models, researchers have proposed a large number of well-established attacks [7–10]. [9] first proposed using adversarial attacks to generate adversarial samples that are recognized correctly by an authorized model while incorrectly by another unauthorized model. However, [9] does not consider unauthorized face recognition models that are not locally available and does not take a transferability improving approach, and the generated sample can not ensure be matched with other samples correctly by the authorized model.

Transfer-based black-box attacks for face models can be borrowed from the attacks for image classification systems due to the similarity of the network structure. In addition, [10] proposes transfer-based attacks for implementing face recognition algorithms. [10] combines [11], [12], and adds dropout units in the middle layer to further enhance the differentiation of the input for each iteration.

In conclusion, none of the above methods can achieve the condition that the generated sample can be correctly recognized by the authorized model while adversarial to the other models. Our approach will focus on improving the transferability of the generated adversarial samples as much as possible under the condition that the success rate of authorized models can be guaranteed.

3. METHODOLOGY

In this section, first, we propose a targeted face image anonymization method that can generate adversarial samples with transferability. Then, we propose constraints on the generation of targeted anonymization samples. Finally, we enhance adversarial transferability to the unauthorized models by proposing a boundary walking strategy.

3.1. Targeted anonymization

Our targeted anonymization method is to modify the original sample to become an adversarial sample to unauthorized models, while the generated sample can be correctly recognized and matched by the authorized model. To generate adversarial samples, we also need at least one model different from the authorized model to be attacked to implement the transfer-based adversarial attack, we call the models the **victim models**. Besides, except for the unauthorized face recognition models, we call the **target models**.

In the process of generating targeted anonymization samples, we have the following objectives:

(1) Guarantee the authorized model can recognize the generated samples correctly. (2) The generated sample can be used as a gallery sample, which means the authorized model can correctly match the generated sample with other face samples of the same person and distinguish it from those belonging to others. (3) The victim and target models have low rates to recognize the generated samples.

The transfer-based attacks generate adversarial samples by attacking the victim models. In general, the lower the output values (confidence scores) of the victim models achieved by the generated adversarial sample, the higher attack success rates on the target models. The white-box attack on the victim model is: find perturbations δ that solves

$$\text{minmize} J\left(f_V\left(x_0+\delta\right), f_V\left(x_0\right)\right), \quad \text{s.t.} \left\|\delta\right\|_\infty<\epsilon \tag{1}$$

where $J(\cdot)$ denotes the cosine similarity. ϵ is the size of perturbation. x_0 is the original sample that needs to be anonymized, and $f_V(\cdot)$ denotes the feature output from the victim model.

To improve the transferability of the generated adversarial samples, we use DI-FGM [12] and MI-FGM [11].

To ensure that the authorized model can recognize correctly, a straightforward approach is to add this condition to Formula (1). So our problem can be solved by finding δ to satisfy

$$\begin{aligned}&\text{minimize} J\left(f_V(x_0+\delta), f_V(x_0)\right)-d\cdot J\left(f_A(x_0+\delta), f_A(x_0)\right)\\&\text{s.t.} \quad \left\|\delta\right\|_\infty<\epsilon\end{aligned} \tag{2}$$

where $f_A(\cdot)$ denotes the authorized model's output. d denotes the weighting of the authorized model in the optimization process. We call this method the baseline method.

However, such a simple approach does not guarantee that the authorized model will recognize correctly when the generated samples are used as gallery samples. If the correct recognition rate of the generated samples, when used as gallery samples, is guaranteed, it can only be done by adding perturbations that ensure that the authorized model is correctly recognized (increase d). This solution causes severe degradation of the adversarial transferability.

3.2. Constraints for generating targeted anonymization samples

Getting the authorized model to recognize correctly does not mean that the generated sample is only recognized as the same person as the original sample; we also want to validate it correctly with other face data. Formula (2) only ensures that the generated sample is recognized as the same user as the original sample, with poor generalization ability.

To solve this problem, we modify the original sample in Formula (2) to the average feature, as shown in Formula (3).

$$\overline{f_A} = \frac{1}{l}\sum_{i=0}^{l} f_A(x_i) \tag{3}$$

where x_i denotes each original sample we have belonging to the user.

In addition, we consider that the decision bound for the average feature is still too broad for our requirement that the generated samples can be matched with other unknown samples. We consider that $J(f_A(x_0+\delta),\overline{f_A}) > J(f_A(x_0),\overline{f_A})$ is a reliable region where the generated samples can be accurately matched with other samples. Because $J(f_A(x_0+\delta),\overline{f_A}) > J(f_A(x_0),\overline{f_A})$ means that the features of the generated samples are more similar to the average features than the features of the original samples. However, a high similarity threshold requires too much perturbation for ensuring that the authorized model satisfies the conditions, resulting in severe degradation of the adversarial transferability to unauthorized models. To this end, we propose a new threshold calculation formula, as shown in Formula (4). By controlling the value of c, we can make a trade-off between the different requirements of authorized and unauthorized models.

$$\begin{aligned} & th_{new} = c \cdot J(\overline{f_A}, f_A(x_0+\delta)) + (1-c)\cdot th_{old} \\ & \text{s.t.} \quad c \in [0,1] \end{aligned} \tag{4}$$

where th_{old} denotes the normal threshold defined by false acceptance rates.

3.3. Boundary walking

For Formula (2), a smaller d ($d \geq 0$) means that the modified perturbations that guide the authorized model to recognize correctly are smaller. However, it is not possible to search for the best value of d in each iteration, which would consume a large number of computational resources. Therefore, we propose boundary walking trying to circumvent the problem of difficulty in setting d.

Boundary walking means that the sample is always located near the decision boundary surface of the authorized model during the optimization process. We used the hyperplane defined by Formula (4) as the boundary surface. Our strategy is to execute the optimization defined in Formula (1) only when the sample is on the correct side of the boundary surface of the authorized model ($J(f_A(x+\delta),\overline{f_A}) \geq th_{new}$), and we execute the Formula (2) when the sample is on the wrong side of the boundary surface of the authorized model ($J(f_A(x+\delta),\overline{f_A}) < th_{new}$). bd selects its value in the way shown in Formula (5).

$$bd = \begin{cases} 0 & \text{if} \quad J(f_A(x+\delta),\overline{f_A}) \geq th_{new} \\ 1 & \text{else} \end{cases} \tag{5}$$

The optimization that uses the boundary walking becomes: find δ that solves

$$\begin{aligned} & \text{minimize} J(f_V(x_0+\delta),\overline{f_V}) - bd \cdot d \cdot J\left(f_A(x_0+\delta),\overline{f_A}\right), \\ & \text{s.t.} \quad \|\delta\|_\infty < \epsilon \end{aligned} \tag{6}$$

As shown in Formula (6), bd can control at each step whether to add perturbations that promote the authorized model to recognize correctly. Then, the boundary walking does not guarantee that the generated samples are located on the correct side, but simply located near the decision boundary. Because we consider that the generated samples located close enough to the boundary are sufficient to achieve our goal. Our method is shown in Algorithm 1.

Algorithm 1 Targeted anonymization with boundary walking

Input: The outputs of the authorized model and victim model $f_A(\cdot)$, $f_V(\cdot)$; the average features of the authorized model and victim model $\overline{f_A}$, $\overline{f_V}$; the original sample x_0; the maximum iterations n, learning rate α, $T(\cdot,p)$ is transformation in [12].
Output: Targeted anonymization samples ta.

```
Let $t = 0$, $x = x_0$, $bd = 0$;
while $t < n$ do
    Diversity input $x_d = T(x,p)$;
    $loss = J(f_V(x_d),\overline{f_V}) - bd \cdot d \cdot J\left(f_A(x),\overline{f_A}\right)$;
    Update $x$ by MI-FGM [11];
    Update $bd$ by Formula (5);
    $t = t + 1$;
end while
return $ta = x$
```

Our approach still combines [11] and [12] for transferability enhancement. Besides, we do not use [12] when we calculate the perturbation that promotes the authorized model to recognize correctly.

4. EXPERIMENT RESULTS

In this section, we first show the settings of our experiments. Next, we show the results of our local evaluation. In the local evaluation, we validate the recognition accuracy of the authorized model when our generated samples are used as gallery samples. Then, we compare our algorithm with the baseline method. Finally, we input the generated adversarial samples into an online FRS API to further evaluate the effectiveness of our method.

4.1. Experiment setting

Our experiments were conducted on the Google Colab cloud server, using a Tesla P100 GPU and PyTorch version is 1.9.0. Our experiments require multiple models for validation. Our models are trained by arcface [13], cosface [14], multimargin [15], with different feature embedding networks. The training dataset we use is CASIA-WebFace [16]. The feature embedding network architectures we use are SE-ResNet50 [13], MobileFaceNet [17], and Attention-56 [18]. To simplify the description, later we directly use the names of these structures to refer to the face recognition models that use them. We use 1000 positive pairs (each pair has at least 5 images belonging to the same person) from the LFW [19] for the experiment. The test is divided into generation and validation stages. In the generation stage, we randomly select one image as the original sample. Then four images of that user including the original sample are randomly selected for computing the average features. In the validation stage, we randomly select one image that has not been selected. We used MTCNN [20] for all the data for alignment with the size of 112x112. Then we define victim models' and the target models' similarity threshold to have a low false acceptance rate (FAR=1e-3).

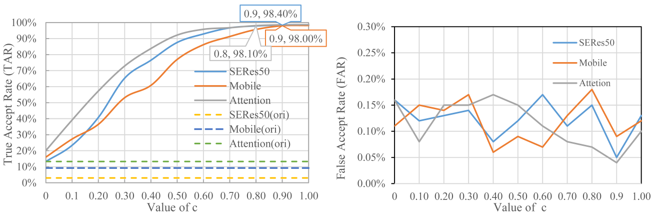
Evaluation Metrics: During the attack, we limit the maximum perturbations that can be tolerated by L_∞. For our attack, there are two main evaluation metrics. (1) The correct recognition rates of the generated data by the target models, the metrics used to evaluate the transferability of the generated adversarial samples. (2) The correct recognition rate of the generated data by the authorized model, which is used to evaluate whether the generated adversarial samples are invalid to the authorized model.

4.2. Local evaluation

For the local experiments, we use SE-ResNet50 as the victim model, other models in turn as the authorized model, and the remaining models as the target model. We set maximum distortion $\epsilon = 8/255$, maximum iterations $n = 20$, decay factor $\mu = 0.6$, learning rate $\alpha = \frac{0.01 \cdot \epsilon}{m}$.

In the local evaluation, first, We show the effects of the c in Formula (4) on the authorized model recognition accuracy. Then, we compare our method with the baseline method.

We show the TAR (True Acceptance Rate) and FAR of the authorized model in the validation phase corresponding to different values of c for different structure networks trained with cosface [14] in Figure 1 (FAR is to calculate the success rate of matching the generated samples with the random sample from other people). In addition, we show the TAR of the authorized model for these networks when using Formula (2) with the normal threshold. The reason we only show the results of the network trained with the cosface loss function is that we found that the networks trained by different loss functions have similar recognition accuracy corresponding to the c.



(a) (b)

Fig. 1. TAR and FAR of different c and network structures. (ori) denotes using $J(f_A(x), f_A(x+\delta)) > th_{old}$ to constrain the range of generated samples

On the one hand, as shown in Figure 1(a), by using Formula (4), we can find the appropriate c value that allows the authorized model to achieve an accuracy of over 98% for the same user (The TAR of SEResNet50, MobileFaceNet, Attention-56 for the original sample are 98.63%, 98.22%, and 98.87%, respectively). While using Formula (2) and the normal threshold, the recognition accuracy would be even lower than 10%. On the other hand, as shown in Figure 1(b), FAR is not significantly affected by changes in c, and the FAR is very close to 1e-3.

We used Formula (2) without any constraints as the baseline method, and a larger value of d to ensure that the generated samples can be recognized correctly by the authorized model. The recognition accuracy of both methods is required to reach near 98%. We show the target model recognition rates with our method and the baseline method in Table 1.

As shown in Table 1, while guaranteeing that the authorized model recognizes correctly, our method can reduce the average recognition accuracy of the target models by at least 40% compared to the baseline method. Our approach sets constraints to precisely achieve our requirements. Besides, our algorithm can compress the perturbations generated to be recognized correctly by the authorized model to a smaller extent than the baseline method.

In addition, we find that whether the target and authorized models are trained with the same loss function does not have a significant effect on the recognition accuracy of the target models, while using the same embedding network causes a significant increase in the recognition accuracy of the target models. Therefore, our approach needs to ensure that the authorized model is not compromised.

Local experiments demonstrate that our method can retain stronger adversarial transferability while guaranteeing the generated samples can be correctly recognized and matched with other samples by the authorized model.

4.3. Online evaluation

We test our method on the commercial API Face++ [21]. Face++ is one of the most popular commercial face recognition platforms. The platform's API can check the likelihood

Table 1. The recognition accuracy of our targeted anonymization method versus the baseline method.

	arcface			multimargin			cosface		
tar. auth.	SE-ResNet	Mobile	Attention	SE-ResNet	Mobile	Attention	SE-ResNet	Mobile	Attention
Mobile(arc.)	0.0%/0.0%[1]	98.6%/98.2%	12.9%/30.2%	0.2%/3.1%	39.3%/52.9%	6.2%/18.8%	2.0%/5.2%	26.0%/46.9%	11.4%/27.3%
Attention(arc.)	0.0%/0.0%	8.2%/19.2%	98.4%/98.3%	0.1%/2.1%	11.2%/24.4%	7.1%/21.3%	1.2%/4.6%	8.1%/18.8%	11.6%/29.7%
SE-ResNet(mul.)	0.0%/0.0%	19.6%/36.8%	15.3%/35.7%	98.5%/98.0%	25.8%/51.2%	12.9%/41.0%	3.0%/12.4%	19.1%/39.5%	21.4%/54.0%
Mobile(mul.)	0.0%/0.0%	21.4%/43.9%	11.2%/25.4%	0.2%/5.3%	98.5%/98.3%	5.1%/16.2%	1.8%/3.1%	19.3%/43.2%	7.1%/30.6%
Attention(mul.)	0.0%/0.0%	14.2%/22.4%	19.2%/36.8%	0.3%/2.9%	13.7%/28.7%	98.2%/98.3%	1.1%/3.1%	8.1%/23.5%	39.2%/66.3%
SE-ResNet(cos.)	0.0%/0.0%	34.6%/49.1%	27.5%/34.6%	7.1%/16.1%	41.8%/47.9%	29.3%/44.8%	98.4%/98.0%	32.6%/44.9%	37.7%/55.1%
Mobile(cos.)	0.0%/0.0%	28.8%/53.7%	14.2%/30.8%	0.4%/4.2%	38.7%/59.2%	6.1%/17.6%	2.7%/9.1%	98.0%/97.8%	11.3%/30.2%
Attention(cos.)	0.0%/0.0%	13.2%/20.3%	19.1%/39.7%	1.3%/2.8%	14.7%/27.4%	23.6%/52.1%	2.1%/6.2%	9.3%/19.5%	98.1%/98.2%
Average[2]	0.0%/0.0%	20.0%/35.1%	17.1%/33.3%	1.4%/5.2%	26.5%/41.7%	12.9%/30.3%	2.0%/6.2%	17.5%/33.8%	20.0%/41.9%

[1] The right side is the result of the baseline method, and the left side is the result of our method.

[2] Average except for the authorized model.

of whether two faces belong to the same person, provide confidence scores and thresholds to evaluate the similarity. We choose the threshold values corresponding to different false acceptance rates (FAR=1e-3,1e-4,1e-5). Due to the speed of the free API test, we used only 200 pairs after MTCNN [20] in our online test.

Compared to the local tests, the FACE++ API is more difficult to execute adversarial attacks. To improve the adversarial transferability of our method, we combine our method with the ensemble attack [22]. We use all the models except the authorized model in the local experiment as the victim models. MobileFaceNet and Attention-56 take turns as the authorized model to represent miniaturized and normal models, and the FACE++ is used as the target model to evaluate transferability. We show the attack results and attack effects when $\epsilon = 8/255$,$\epsilon = 10/255$ respectively. Other parameters are consistent with the local evaluation.

Table 2. Recognition rates for online API with the multiple victim models of our method

Authorized models	FAR			Confidence
	1e-3	1e-4	1e-5	
clean[1]	100.0%	100.0%	99.5%	88.24
-[2](8/255)	18.5%	6.0%	1.5%	49.53
Mobile (8/255)	32.0%	16.0%	5.0%	55.21
Attention (8/255)	37.0%	20.5%	10.0%	58.25
-[2](10/255)	6.0%	1.0%	0.0%	41.1
Mobile (10/255)	12.5%	3.0%	1.0%	47.51
Attention (10/255)	17.5%	5.5%	3.0%	50.23

[1] clean denotes the original samples.

[2] - denotes the adversarial samples generated by transfer-based attack.

As shown in Table 2, generated samples make the confidence level of FACE++ drop significantly. When $\epsilon = 8/255$, our algorithm can reduce the recognition success rate by more than 60% (FAR=1e-3). When $\epsilon = 10/255$, it can reduce the recognition accuracy by more than 80% (FAR=1e-3). Increasing the amount of perturbation can significantly reduce the accuracy of recognition. In addition, the generated samples are guaranteed to be correctly recognized by the authorized model, reducing its adversarial transferability.

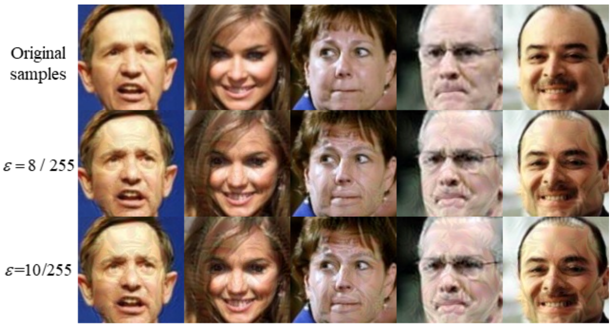


Fig. 2. Targeted anonymization results of our method.

The samples we generate to attack the commercial API are shown in Figure 2. The samples generated by our method perform transfer-based attacks on FACE++ under conditions of acceptable perturbations and can be correctly recognized by the authorized model.

The online evaluation demonstrates that our method can achieve anonymization with a high probability for the current state-of-the-art face recognition system while ensuring that the generated samples can be recognized by human eyes and the authorized model.

5. CONCLUSION

In this paper, we propose a face recognition anonymization scenario when we want the face images can only be recognized by the authorized model and human eyes. For this scenario requirement, we propose constraints for generating samples and the boundary walking strategy. The local evaluations demonstrate that our method is effective in generating adversarial samples that are invalid for the authorized model but valid for other models. Also, compared to the baseline approach, our algorithm can significantly reduce the recognition success rate of target models. Finally, we attack the FACE++ API to verify the effectiveness of our method in real scenarios.

6. ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of China under Grant 62102379 and 61902262, in part by the National Key Research and Development Program of China under Grant 2018YFB0804103.

7. REFERENCES

[1] Carman Neustaedter, Saul Greenberg, and Michael Boyle, "Blur filtration fails to preserve privacy for home-based video conferencing," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 13, no. 1, pp. 1–36, 2006.

[2] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker, "Model-based face de-identification," in *2006 Conference on computer vision and pattern recognition workshop (CVPRW'06)*. IEEE, 2006, pp. 161–161.

[3] Vahid Mirjalili, Sebastian Raschka, and Arun Ross, "Privacynet: semi-adversarial networks for multi-attribute face privacy," *IEEE Transactions on Image Processing*, vol. 29, pp. 9400–9412, 2020.

[4] Jiacheng Lin, Yang Li, and Guanci Yang, "Fpgan: Face de-identification method with generative adversarial networks for social robots," *Neural Networks*, vol. 133, pp. 132–147, 2021.

[5] Jiaming Zhang, Jitao Sang, Xian Zhao, Xiaowen Huang, Yanfeng Sun, and Yongli Hu, "Adversarial privacy-preserving filter," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1423–1431.

[6] Xiao Yang, Yinpeng Dong, Tianyu Pang, Jun Zhu, and Hang Su, "Towards privacy protection by generating adversarial identity masks," *arXiv preprint arXiv:2003.06814*, 2020.

[7] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter, "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*, 2016, pp. 1528–1540.

[8] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7714–7722.

[9] Hyun Kwon, Ohmin Kwon, Hyunsoo Yoon, and Ki-Woong Park, "Face friend-safe adversarial example on face recognition system," in *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2019, pp. 547–551.

[10] Yaoyao Zhong and Weihong Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1452–1466, 2020.

[11] 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*, 2018, pp. 9185–9193.

[12] 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*, 2019, pp. 2730–2739.

[13] 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*, 2019, pp. 4690–4699.

[14] 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*, 2018, pp. 5265–5274.

[15] Bing Cao, Nannan Wang, Xinbo Gao, Jie Li, and Zhifeng Li, "Multi-margin based decorrelation learning for heterogeneous face recognition," in *IJCAI*, 2019.

[16] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[17] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[18] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[19] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "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.

[20] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[21] "Face++ research toolkit," Available:,`http://www.faceplusplus.com.cn`, 2021.

[22] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.