RAF: Recursive Adversarial Attacks on Face Recognition Using Extremely Limited Queries

Keshav Kasichainula
Computer Science Department
University Of Houston
Houston, USA
kkasichainula@uh.edu

Hadi Mansourifar
Computer Science Department
University Of Houston
Houston, USA
hmansourifar@uh.edu

Weidong Shi
Computer Science Department
University Of Houston
Houston, USA
wshi3@uh.edu

Abstract—Recent successful adversarial attacks on face recognition show that, despite the remarkable progress of face recognition models, they are still far behind human intelligence for perception and recognition. It reveals the vulnerability of deep convolutional neural networks (CNNs) as a state-of-the-art building block for face recognition models against adversarial examples, which can cause certain consequences for secure systems. Gradient-based adversarial attacks have been widely studied and proved successful against face recognition models. However, finding the optimized perturbation per each face needs to submit a significant number of queries to the target model. In this paper, we propose a recursive adversarial attack on face recognition using automatic face warping, which needs an extremely limited number of queries to fool the target model. Instead of random face warping procedure, the warping functions are applied on specific detected regions of face like eyebrows, nose, lips, etc. We evaluate the robustness of the proposed method in the decision-based black-box attack setting, where the attackers have no access to the model parameters and gradients, but the target model provides hard-label predictions and confidence scores.

Index Terms—Adversarial Attack, Face Recognition, Machine Learning Security

I. INTRODUCTION

Face recognition [10]–[12] is one of the most well-known computer vision tasks which deep CNNs substantially improve. Face recognition includes two different tasks: face verification and face identification. Face verification compares a pair of face images to investigate the same identity. In contrast, face identification is a classification task to assign a label as an identity to a face image. Face embeddings are state-of-the-art and are widely used for both tasks as deep face features, which make minimum intra-class and maximum inter-class variances among face classes. The remarkable performance of face embeddings has made face recognition very popular for identity authentication in a wide range of security-sensitive applications from financial sectors to criminal identification.

However, deep CNNs proved to be vulnerable against adversarial instances [13]–[16]. Adversarial inputs [17] are defined as maliciously generated indistinguishable instances for human eyes by adding small perturbations [18]. The adversarial attacks against face recognition are often divided into two types: dodging attacks that enable attackers to evade being recognized and impersonation attacks in which the attacker is recognized as another individual. For example, one traditional way of adding perturbations to the face is to wear glasses which can fool the target model for null or wrong recognition. The adversarial attacks are evaluated in two settings: White box and Black box. In a white-box setting, the attackers know the architecture and hyper-parameters of the target model, so gradient-based methods can directly optimize the loss function. Needless to say, this scenario is not compatible with real-world cases since the attackers cannot get access to the model details. The black-box setting is more realistic since no internal model information is known to the attackers except the model’s output, including hard-label prediction and confidence score. Designing such adversarial attacks based on perturbation proved successful but with the main problem: finding the optimal perturbations needs many queries. In recent years, geometrically-perturbed face images have successfully fooled the FR [8], [9]. However, these approaches face a trade-off: the adversarial face images are low quality, or a huge number of queries are needed to reach a high-quality adversarial instance. In this paper, we propose a Recursive Adversarial Attack on FR (RAF) using smart face warping to achieve acceptable qualities with an extremely limited number of queries. First, we use smart face warping to decompose the face regions into
areas with a minimum overlap which helps to reduce the size of state space significantly. Second, we form a Depth First Search (DFS) tree with a different warping function as its nodes to search the state space. As a result, a different warping function is applied to the face image as we traverse the DFS tree recursively. Third, the adversarial instance is submitted to the target model to investigate if the dodging or impersonating goal has been reached, given the returned label from the model. The proposed attack has three main characteristics: (i) The adversarial instances are not generated by adding perturbations. (ii) The attacker can reach acceptable dodging or impersonating instances with extremely limited attempts (less than seven queries), (iii) The adversarial threat can not be removed by pre-processing. To demonstrate the significant success of our proposed attack, we use Amazon Rekognition and different VGGFace models. Amazon Rekognition is a cloud-based software as a service (SaaS) computer vision platform introduced in 2016. The technology has been used by law enforcement agencies and was reportedly pitched to Immigration and Customs Enforcement (ICE) in the U.S [19]. Figure 1 shows the returned results from Rekognition. Our contributions are as follows.

1) We propose the first adversarial attack on face recognition using smart face warping.
2) We show that the state of the art cloud-based paid face recognition service is significantly vulnerable against proposed attack.
3) Our experiments show that, the attacker can reach the dodging or impersonation instances after very limited queries.

Table I summarizes the previous works in the field of adversarial attacks on face recognition systems.

The rest of paper is organized as follows. Section 2 talks about the adversarial face warping attack. Section 3 shows the experiments and finally section 4 concludes the paper.

II. ADVERSARIAL FACE WARPING ATTACK

This section presents the proposed black-box attack setting and evolutionary attack method against a face recognition model.

1) https://aws.amazon.com/rekognition/

A. Attack Setting

A target face recognition model is denoted by \( f(x) : \chi \rightarrow \gamma \) (\( \chi \subseteq \mathbb{R}^n \)), where in face identification task the input image \( x \) is compared with a gallery set of face images, and then classifier assigns the \( x \) a specific identity such that, \( \gamma = 1, 2, ..., K \) where \( K \) is the number of identities. Supposing a target face image \( x \) and correct label \( \gamma \), the goal of attacker is to warp \( x \) to create an adversarial face image \( x^w \) and get label \( \gamma^w \) from the model such that \( \gamma^w \neq \gamma \) with minimum changes comparing to \( x \). It can be obtained by solving a constrained optimization problem as follows.

\[
\min_{x^w} \Delta(x, x^w), \ s.t. \ \gamma \neq \gamma^w \tag{1}
\]

We use the \( L_2 \) distance to calculate the \( \Delta \). Equation (2), the constrained problem can be equivalently reformulated as the following unconstrained optimization problem.

\[
\min_{x^w} \mathcal{L}(x^w) = \Delta(x, x^w) + \delta(\gamma \neq \gamma^w) \tag{2}
\]

where \( \delta(x, x^w) = 0 \) if \( \gamma \neq \gamma^w \), otherwise \( \delta(x, x^w) = \infty \). As a result, the adversarial instance \( x^w \) is supposed to be obtained with the minimum required perturbation by optimizing Eq. (3).

1) Dodging: This attack generates an adversarial image that is not recognized as an identity in the training data.

\[
\exists x^w \ni, \ s.t. \ \forall \gamma (\gamma_i \neq \gamma^w) \tag{3}
\]

2) Impersonation: This attack corresponds to generating an adversarial image that is recognized as a wrong identity in the training data.

\[
\forall x_i \exists x^w_j (\gamma_i = \gamma^w_j), \ s.t. \ i \neq j \tag{4}
\]

B. Recursive Adversarial Attack

One of the most significant deficiencies of old adversarial attacks is that the output is insensitive to small input perturbations. That’s why the attacker needs many queries to reach a successful adversarial instance, and the gradient estimation methods cannot be directly used. Some methods [20], [21] successfully reformulated the discontinuous optimization problem in Eq. (2) as some continuous optimization problems and used
gradient estimation methods for optimization. But they need to calculate the distance of a point to the decision boundary or estimate the predicted probability by the hard-label outputs, which are less efficient. This paper uses recursive optimization to find the minimum warping required to fool the target model. Recursive optimization is an evolutionary computation framework that can be used to solve high-dimensional optimization problems via a 'divide-and-conquer' mechanism, where the main challenge lies in problem decomposition.

1) Definition 1: \( f(x) \) is globally decomposable if there exists a partition \( \{x_C, x_{U1}, x_{U2}\} \) such that, for every partial assignment \( \rho_C f_{\rho_C}(x_{U1}, x_{U2}) = f_{\rho_C}(x_{U1}) + f_{\rho_C}(x_{U2}) \). According to Definition 1 [22], face warping problem can be decomposed into a set of sub-functions like smile, raise eyebrow, nose stretch, etc., since each of the sub-functions are independent from each other with minimum overlap in terms of impacted face regions. As a result, (i) face warping functions enable the attacker to avoid random perturbations by applying smart face warping. (ii) The state space can be decomposed to non-overlapping regions by applying sub-function recursively.

Figure 3 illustrates the decomposition of the state space into smaller regions by divide and conquer strategy [22]. Each node represents a different warping region to be processed by a different warping function with different operation and scale.

2) Proposition 1: If at each level, RAF chooses \( x_C \subseteq x \) of size \( |x_C| = d \) such that, for each selected value \( \rho_C \), the simplified function \( f_{\rho_C}(x_U) \) locally decomposes into \( k > 1 \) independent sub-function \( \{f_i(x_{U1})\} \) with equal-sized domains \( x_{U1} \), then the time complexity of RAF is \( O\left(\frac{n}{d}\xi(d)\log_2\left(\frac{n}{d}\right)\right) \) with following proof [22]

Proof. Assuming that it is of the same order as \( n \), the recurrence relation for RDIS is

\[
T(n) = O(n) + \xi(d)[g(d) + O(m) + O(n)] + O(d\log^2(n)) + kT\left(\frac{n-d}{k}\right)
\]

which can be simplified to

\[
T(n) = \xi(d)[kT\left(\frac{n}{k}\right)] + O(n) + O(n)
\]

Noting that the recursion halts at \( T(d) \), the solution to the above recurrence relation is then

\[
T(n) = c_1k\xi(d)\log_2\left(\frac{n}{d}\right) + c_2n\sum_{r=0}^{\log_2\left(\frac{n}{d}\right)}\xi(d)^r
\]

which is

\[
O(k\xi(d)\log_2\left(\frac{n}{d}\right)) = O\left(\frac{n}{d}\xi(d)\log_2\left(\frac{n}{d}\right)\right)
\]

Algorithm 1 shows the required steps for a recursive adversarial attack on face recognition, where \( WF_i \) represents the active warping function that applies the automatic warping on the given image denoted by \( x \). First, we create a warping function set containing a sequence of warping functions. We need to traverse a state space of warping functions and apply them one by one until finding an adversarial instance. Since we want to find an adversarial instance in extremely limited queries, we set a boundary denoted by \( l \). Also, the \( id \) is used to check the original target identity with the returned identity from the target model \( (rid) \). The recursive algorithm stops in case \( id \neq rid \), which means a successful impersonation attack.
but if $rid = Null$, it's a successful dodge attack. The other condition to stop is the number of queries.

C. DFS Search Tree

Although the recursive adversarial search is supposed to traverse all the warping functions, the order of nodes in the DFS tree can impact some faces’ required number of queries. That is why we need to investigate which one of the warping functions has more impact on adversarial attacks. Our empirical experiments show that three warping functions, including Smile, Raise Eyebrow, and Stretch Nose have the highest impacts on high-quality warped faces. Thus, we decided to keep the DFS tree limited to three mentioned functions, as shown in Figure 2.

D. Hybrid Adversarial Faces

Although many adversarial instances can be generated using only one warping function, sometimes more than one warping function is needed to fool the target model. Such face images are called hybrid adversarial instances, as shown in Figure 4.

III. EXPERIMENTS

This section presents the experimental setups, results, and discussions related to proposed adversarial attacks on face recognition. Figure 5 is an example of an impersonation attack on a famous celebrity image.

A. Datasets

For the sake of experiments, we used the following datasets. **Collected Dataset:** We collected 101 casual celebrity faces from the web, which mostly contains frontal celebrity faces. In addition, we used CelebA Dataset [23] and CASIA-WebFace Dataset [24].

B. Target Models

In addition to creating FR models using VGG16, Resnet50 [25] and Senet50 [26], we also used following target models. **AWS Rekognition** A cloud-based software as a service computer vision platform. Figure 5 shows a sample result using AWS Rekognition. **VGGFace** A series of models developed for FR by Visual Geometry Group (VGG).

C. Attack Evaluation with Extremely Limited Queries

To evaluate the proposed method, we selected the most challenging scenario in which the attacker has only three chances to fool the target model. In other words, we assumed that the target model accepts only three queries from each authorized user at a specific time. The DFS tree looks like the highlighted path by orange arrows in Figure 2. Our experimental results show that the proposed adversarial attack using the automatic face warping reaches a high success rate with minimal queries. We can summarize the limited query results as follows. (i) In the case of dodge attack in the collected dataset, most successful attacks find the adversarial instance with only one query. (ii) In the case of dodge attack in CelebA and CASIA Web datasets, most successful attacks find the adversarial instance with two queries. (iii) In the case of impersonation attacks in the collected dataset, most successful attacks find the adversarial instance with three queries. (iv) In the case of impersonation attack in CelebA dataset, most successful attacks find the adversarial instance with two queries. (v) In all tested datasets, finding the impersonation instance is very challenging compared to dodge instances.

D. Adversarial Instance Samples

This section presents some successful adversarial instances in both dodge and impersonation categories created by our proposed method. Figure 5 shows a print screen of the AWS Rekognition page with an original face correctly recognized and a wrongly recognized warped face which was obtained by two queries (Raise eyebrow, Scale=0.2). Also, Part (A) of Figure 6 shows some failed and successful adversarial instances, and the details can be found in Table III. Note that parts (a,b,c) are selected from the collected dataset while parts (d,e) are from CelebA and CASIA Web datasets, respectively.

E. Adversarial Instance Samples

This section presents some successful adversarial instances in both dodge and impersonation categories created by our proposed method. Figure 5 shows a print screen of the AWS Rekognition page with an original face correctly recognized and a wrongly recognized warped face which was obtained by two queries (Raise eyebrow, Scale=0.2). Also, Part (A) of Figure 6 shows some failed and successful adversarial instances, and the details can be found in Table III. Note that
TABLE II
COMPARING DIFFERENT TARGET FR MODELS USING RAF. THE NUMBERS BELOW INDICATE THE NUMBER OF DODGE AND IMPERSONATING RESULTS RECEIVED FROM ALL THE WARPED IMAGES SUBMITTED TO THE MODELS IN LESS THAN 4 QUERIES. RE - RAISED EYEBROWS AND SN - STRETCHED NOSE.

|                  | Dodge Attack | Impersonation Attack |
|------------------|--------------|----------------------|
|                  | CelebA       | Web Casia            | Collected |
|                  | Smile (0.1)  | RE (0.2)             | SN (0.3)  |
| AWS Recognition  | 47           | 16                   | 184       |
| VGG 16           | 47           | 56                   | 48        |
| Resnet50 [25]    | 24           | 10                   | 19        |
| Senet50 [26]     | 16           | 43                   | 40        |

TABLE III
DETAIL DESCRIPTION OF ADVERSARIAL INSTANCES PRESENTED IN PART (A) OF FIGURE 6. RE DENOTES RAISED EYEBROWS, SN DENOTES STRETCHED NOSE, D DENOTES DODGE ATTACK, I DENOTES IMPERSONATING ATTACK AND N DENOTES NO ATTACK.

| Figure | Attack Type | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|--------|-------------|-----|-----|-----|-----|-----|-----|-----|-----|
| (a)    |             | N   | N   | D   | N   | D   | I   |     |     |
| (b)    |             | I   | D   | D   | N   | D   | D   | I   |     |
| (c)    |             | N   | N   | N   | N   | D   | I   | D   | D   |
| (d)    |             | N   | N   | D   | N   | D   | D   | N   | D   |
| (e)    |             |     |     |     |     |     |     |     |     |

Fig. 6. (A) Adversarial instances created only by one warping function. Hybrid attack with (B) 5 queries and (C) 6 queries : The successful adversarial instances found in fifth and sixth queries respectively (the last columns in part (B),(C).

parts (a,b,c) are selected from the collected dataset while parts (d,e) are from CelebA and CASIA Web datasets, respectively.

F. Hybrid Adversarial Instances

Although we selected the most challenging scene to find the adversarial instances with less than four queries, we also studied the warped faces which need more than three queries to fool the target model. Such warped faces are hybrid adversarial instances since more than one warping function contributes to warping them. Figure 4 shows a case study that took five queries to reach a successful dodge attack.

• Creating Adversarial instances in 4 Queries

In the case of four queries, we perform the following functions on the images. Raised Eyebrows(0.1), Raised Eyebrows(0.2), Raised Eyebrows(0.2), Raised Eyebrows(0.3) + Smile(0.1), Raised Eyebrows(0.3) + Smile(0.1), Raised Eyebrows(0.3) + Smile(0.1) + Stretched Nose(0.1). Moreover, the generated images are checked after each instance to see if they qualify as adversarial. We performed this experiment on all the datasets (collected, celebA, Web Casia) and on all the models (AWS, VGG16, Resnet50, Senet50).

• Creating Adversarial instances in 5 Queries

In the case of 5 queries, we perform the following functions on the images. Eyebrows(0.1), Raised Eyebrows(0.2), Raised Eyebrows(0.2), Raised Eyebrows(0.3) + Smile(0.1), Raised Eyebrows(0.3) + Smile(0.1) + Stretched Nose(0.1). Moreover, the generated images after each instance are checked to see if they qualify as an adversarial instance or not. We performed this experiment for all the datasets (collected, celebA, Web Casia) and also on all the models (AWS, VGG16, Resnet50, Senet50).

• Creating Adversarial instances in 6 Queries

In the case of 6 queries, we perform the following functions on the images. Raised Eyebrows(0.1), Raised Eyebrows(0.2), Raised Eyebrows(0.2), Raised Eyebrows(0.3) + Smile(0.1), Raised Eyebrows(0.3) + Smile(0.1) + Stretched Nose(0.1), Raised Eye-
brows(0.3) + Smile(0.2) + Stretched Nose(0.1) And the generated images after each instance are checked to see if they qualify as an adversarial instance or not. We performed this experiment on all the datasets (collected, celebA, Web Caisia) and on all the models (AWS, VGG16, Resnet50, Senet50).

The results can be found in the table IV. They are categorized by the number of queries and successful images for each dataset. Please note that for VGGFace models like the VGG16, Resnet50, Senet50, we selected the confidence threshold to be 60% and above. Any confidence score below that was discarded.

### IV. Conclusion

This paper proposes a new adversarial attack on face recognition models with extremely limited queries. We used automatic face warping in a recursive adversarial context to target the commercial and public face recognition-retrieval models. Our experiments showed that the proposed attack could fool the online FR models with less than four queries.

### References

[1] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessory 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.

[2] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. 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.

[3] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, “Invisible mask: Practical attacks on face recognition with infrared,” *arXiv preprint arXiv:1803.04683*, 2018.

[4] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang, “Generating adversarial examples by makeup attacks on face recognition,” in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2516–2520.

[5] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, “Adversarial light projection attacks on face recognition systems: A feasibility study,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 814–815.

[6] Z. Xiao, X. Gao, C. Fu, Y. Dong, W. Gao, X. Zhang, J. Zhou, and J. Zhu, “Improving transferability of adversarial patches on face recognition with generative models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11845–11854.

[7] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, “Unravelling robustness of deep learning based face recognition against adversarial attacks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[8] R. Wang, F. Juefei-Xu, Q. Guo, Y. Huang, X. Xie, L. Ma, and Y. Liu, “Amora: Black-box adversarial morphing attack,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1376–1385.

[9] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, “Fast geometrically-perturbed adversarial faces,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1979–1988.

[10] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arfaced: 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.

[12] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, “Learning meta face recognition in unseen domains,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6163–6172.

[13] E. Radiya-Dixit and F. Tramèr, “Data poisoning won’t save you from facial recognition,” *arXiv preprint arXiv:2106.14851*, 2021.

[14] M. Fredriksson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[15] H. Mansourifard and W. Shi, “Vulnerability of face recognition systems against composite face reconstruction attack,” *arXiv preprint arXiv:2009.02286*, 2020.

[16] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, “Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system,” *JSI Transactions on Internet and Information Systems (TIIS)*, vol. 15, no. 3, pp. 1100–1118, 2021.

[17] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, “When nas meets robustness: In search of robust architectures against adversarial attacks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 631–640.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[19] E. Culliford, “From Clubhouse to Twitter Spaces, social media grapples with live audio moderation,” https://www.cnbc.com/2020/06/10/amazon-bans-police-use-of-facial-recognition-technology-for-one-year.html, 2021.

[20] R. Maheshwary, S. Maheshwary, and V. Pudi, “Generating natural language attacks in a hard label black box setting,” in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[21] P. Saadatpanah, A. Shafahi, and T. Goldstein, “Adversarial attacks on copyright detection systems,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 8307–8315.

[22] A. L. Friesen and P. Domingos, “Recursive decomposition for nonconvex optimization,” in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, “Large-scale celebfaces attributes (celeba) dataset,” *Retrieved August*, vol. 15, no. 2018, p. 11, 2018.

[24] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *arXiv preprint arXiv:1411.7923*, 2014.

[25] D. Theckedath and R. Sedarikar, “Detecting affect states using vgg16, resnet50 and se-resnet50 networks,” *SN Computer Science*, vol. 1, no. 2, pp. 1–7, 2020.

[26] W. Luo, X. Yang, X. Mo, Y. Lu, L. S. Davis, J. Li, J. Yang, and S.-N. Lim, “Cross-x learning for fine-grained visual categorization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8242–8251.