Pattern Recognition Letters 140 (2020) 222–229

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Cross-resolution face recognition adversarial attacks

Fabio Valerio Massoli*, Fabrizio Falchi, Giuseppe Amato

*ISTI-CNR, via G. Moruzzi 1, Pisa 56124, Italy*

ARTICLE INFO

*Article history:*
Received 17 July 2020
Revised 21 September 2020
Accepted 14 October 2020
Available online 15 October 2020

*Keywords:*
Deep learning
Face recognition
Adversarial attacks
Face identification
Adversarial biometrics

ABSTRACT

Face Recognition is among the best examples of computer vision problems where the supremacy of deep learning techniques compared to standard ones is undeniable. Unfortunately, it has been shown that they are vulnerable to adversarial examples - input images to which a human imperceptible perturbation is added to lead a learning model to output a wrong prediction.

Moreover, in applications such as biometric systems and forensics, cross-resolution scenarios are easily met with a non-negligible impact on the recognition performance and adversary's success. Despite the existence of such vulnerabilities set a harsh limit to the spread of deep learning-based face recognition systems to real-world applications, a comprehensive analysis of their behavior when threatened in a cross-resolution setting is missing in the literature.

In this context, we posit our study, where we harness several of the strongest adversarial attacks against deep learning-based face recognition systems considering the cross-resolution domain. To craft adversarial instances, we exploit attacks based on three different metrics, i.e., $L_1$, $L_2$, and $L_\infty$, and we study the resilience of the models across resolutions. We then evaluate the performance of the systems against the face identification protocol, open- and close-set.

In our study, we find that the deep representation attacks represents a much dangerous menace to a face recognition system than the ones based on the classification output independently from the used metric. Furthermore, we notice that the input image's resolution has a non-negligible impact on an adversary's success in deceiving a learning model. Finally, by comparing the performance of the threatened networks under analysis, we show how they can benefit from a cross-resolution training approach in terms of resilience to adversarial attacks.

## 1. Introduction

Face Recognition [1,2] (FR) represents one of the most astonishing applications of Neural Networks (NNs), especially considering Deep Convolutional Neural Networks (DCNNs), that ultimately overcame standard computer vision techniques such as Gabor-Fisher [3] and local binary patterns [4]. The study of such a problem began in the early 90s when [5] proposed the Eigenfaces approach, and it only required two decades for Deep Learning (DL) approaches to start to dominate the field reaching recognition performance up to 99.80% [1], thus overcoming human ability. DL-based FR systems do not exploit the output of a classifier directly. Instead, they leverage the representation power [6] of the learning models to extract face descriptors, i.e., multidimensional vectors, also called deep features or deep representations, to fulfill the recognition task.

Although FR systems obtain very high performance when trained with datasets comprising images acquired under controlled conditions, e.g., high-resolution, they suffer a drastic drop in reliability when tested against cross-resolution (CR) scenarios [7] that naturally arise, for example, in surveillance applications [8–10]. To counteract such a weakness, Ekenel and Sankur [11] and Luo et al. [12] proposed approaches that were not based on NNs. Instead, only recently such a problem has been tackled in the DL field [13,14].

To make the situation even worse, recently [15,16] showed that DL models are vulnerable to the so-called *adversarial* examples - images to which a specific amount of noise, undetectable to humans, is added to induce a NN to output a wrong prediction. Unfortunately, the ability of an insightful adversary to jeopardize these learning models, considering both the digital [17–21] and physical [22,23] domains, represents a significant concern in security-related applications such as DL-based biometrics systems [24] and forensics [25]. Thus, limiting their adoption in these fields.

* Corresponding author.
*E-mail address:* fabio.massoli@isti.cnr.it (F.V. Massoli).

https://doi.org/10.1016/j.patrec.2020.10.008
0167-8655/© 2020 Elsevier B.V. All rights reserved.

In this context, we posit our contribution that we summarize as follows: i) we threaten two DCNNs by exploiting adversarial attacks based on three different metrics, i.e., $L_1$, $L_2$, and $L_\infty$; ii) we generate attacks not only towards a classification objective but also against a similarity one. Indeed, FR systems typically do not exploit a DCNN classification output. Instead, they leverage the ability of NNs to generate discriminative deep representations among which a similarity criterion is evaluated to fulfill the recognition task; iii) we conduct the attacks in a cross-resolution domain, thus emulating a real-world scenario for an FR system; iv) we analyze the success rates of the various attacks across resolutions, studying if a DL model can benefit from a cross-resolution training procedure in terms of robustness to adversarial attacks; v) we analyze the robustness of the models through the face identification protocol [26] considering both the open- and close-set settings.

The rest of the paper is structured as follows. In Section 2, we briefly present some related works, while in Section 3, we describe the attacks algorithms we use. Subsequently, in Section 4, we explain our experimental procedure and the dataset we use, while in Section 5, we present the results from the experimental campaign. Finally, in Section 6, we report our conclusions.

## 2. Related works

To the best of our knowledge, this is the first work that tackles the problem of adversarial attacks against FR systems in a CR scenario. For such a reason, in what follows, we briefly cite a few articles related to the topics of the cross-resolution FR and adversarial attacks against an FR system.

### 2.1. Cross-resolution face recognition

CR scenarios are met whenever images at different resolutions have to be matched. Such a situation typically happens, for example, in biometric and forensics applications. Super-Resolution (SR) techniques are among the most studied solutions to such a problem, and Singh et al. [27] proposed to synthesize high-resolution faces from low-resolution ones by employing a multi-level sparse representation of the given inputs. Zangeneh et al. [28] formulated a mapping of the low- and the high-resolution images to a common space by leveraging a DL architecture made by two distinct branches, one for each image. Luo et al. [12] exploited the dictionary learning approach based on learning multiple dictionaries, each being associated with a resolution. The most comprehensive study and widely tested method to improve an FR system's performance in a CR scenario was recently proposed by Massoli et al. [13]. In their work, the authors formulated a training procedure to fine-tune a state-of-the-art model to the CR domain. They tested their models on several benchmark datasets by showing their superior performance compared to the results available in the literature.

### 2.2. Face recognition adversarial attacks

As we mentioned at the beginning of this section, we are the first to study adversarial attacks in a cross-resolution domain. Due to the lack of papers than can be directly compared to our study, in what follows we only briefly cite a few articles concerning adversarial attacks against FR systems. Sharif et al. [22] demonstrated the feasibility and effectiveness of physical attacks by impersonating other identities using eyeglass frames with a malicious texture. Zhong and Deng [29] observed the superior transferability properties of feature-based attacks compared to label-based ones. Moreover, they proposed a drop-out method for DCNNs to enhance further the transferability of the attacks. Song et al. [18] proposed a three-player GAN architecture that leveraged a face recognition network as the third player in the competition between generator and discriminator. Dong et al. [17] successfully performed black-box attacks on FR models and demonstrated their effectiveness in a real-world deployed system.

Face recognition is a sensitive topic since it usually involves persons' privacy. Several techniques have been proposed in the literature to protect people's identities, such as the Fawkes algorithm [30]. The goal of such a technique is to modify a user image so that a face model trained on the manipulated images will not recognize genuine images of the original subject. However, such an approach is based on a different principle than the adversary-defender arm race one, thus requiring a completely different analysis to the one we present in our work. For such a reason, we do not consider it in our analysis.

## 3. Adversarial attacks

### 3.1. Carlini and wagner - CW

Carlini and Wagner [31] formulated one of the strongest currently available attacks. The CW-$L_2$ attack is formalized as:

$\min\ c \cdot f(\frac{1}{2}\tanh(\mathbf{w})+1)+ \parallel \frac{1}{2}(\tanh(\mathbf{w})+1)-\mathbf{x} \parallel_2^2$, where $f(\,\cdot\,)$ is the objective function, $\mathbf{x}$ is the input image, $\mathbf{w}$ is the adversarial example in the tanh space, and $c$ is a positive constant which value is set by exploiting a binary search procedure.

### 3.2. Elastic net attack - EAD

The EAD Attack [32], leverages the elastic-net regularization which is a well known technique in solving high-dimensional feature selection problems [33]. It is based on the objective proposed in Carlini and Wagner [31] and it conceives the CW-$L_2$ attack as a special case. EAD is formulated as:

$\min_{\mathbf{x}}\ c \cdot f(\mathbf{x}, t) + \beta \parallel \mathbf{x} - \mathbf{x}_0 \parallel_1 + \parallel \mathbf{x} - \mathbf{x}_0 \parallel_2^2$, where $f(\,\cdot\,)$ is the objective as in the CW-$L_2$ attack, $t$ is the target class, $\mathbf{x}_0$ is the input image, $t$ is the target label, $\mathbf{x}$ is the adversarial instance, $c$ is a parameter found by binary search, and $\beta$ represents the weight of the $L_1$ penalty term.

### 3.3. Jacobian saliency map attack - JSMA

The JSMA [34] attack exploits an "input-perturbation-to-output" mapping. Differently from the backpropagation-based attacks, JSMA leverages the model derivative concerning the classification output rather than the derivative of the loss function. The attack is formalized as: $\arg\min_{\delta_{\mathbf{x}}} \parallel \delta_{\mathbf{x}} \parallel \quad \text{s.t.}\ \mathbf{F}(\mathbf{X}+\delta_{\mathbf{x}}) = \mathbf{Y}^*$, where $\mathbf{F}$ is the function learned by the DNN, $\mathbf{X}$ and $\mathbf{Y^*}$ are the input and output of the model, respectively, and $\delta_{\mathbf{x}}$ is the adversarial perturbation defined upon the evaluation of the model input saliency map.

### 3.4. Deep representations attacks - DR

Differently from the previously mentioned attacks, the Deep Representations [35] attack focuses on the manipulation of image features. It is formulated as an optimization problem which aims at finding the closest perturbed image, to the original one, whose descriptor is as close as possible to the one of a target image named the "guide image". Specifically, the adversarials crafting procedure is the following: $\mathbf{I}_\alpha = \arg\min_{\mathbf{I}} \parallel \phi_k(\mathbf{I}) - \phi_k(\mathbf{I}_g) \parallel_2^2$; subject to $\parallel \mathbf{I} - \mathbf{I}_s \parallel_\infty < \delta$, where $\phi(\,\cdot\,)_k$ is the descriptor extracted at layer $k$ of the threatened model, $\mathbf{I}_s$ and $\mathbf{I}_g$ are the source and target images, respectively, $\mathbf{I}_\alpha$ is the adversarial example, and $\delta$ is he maximum allowed perturbation in terms of the $L_\infty$ norm.

### 3.5. Projected gradient descent - PGD

The PGD attack, was proposed by Madry et al. [36]. It applies the FGSM [37] attack multiple times with small step size. It is formalized as:

$$\mathbf{x}_{N+1}^{adv} = Clip_{\mathbf{x},\epsilon}\left\{\mathbf{x}_N^{adv} + \alpha \cdot \mathrm{sign}(\nabla_{\mathbf{x}} J_\theta(\mathbf{x}_N^{adv}, y))\right\}, \quad (1)$$

where the $Clip(\ \cdot\ )$ function clips the values of the pixels to the allowed range, and $\alpha$ is the step size. The iteration starts from an acceptable random perturbation of the input $\mathbf{x}_{N=0}$.

## 4. Experimental approach

### 4.1. Dataset and models

In our experiments, we use two datasets: VGGFace2 [38] and SCface [39].

The *VGGFace2* [38] dataset contains a training set made by ~ 2.9 M images shared among 8631 identities. To construct the gallery and the queries, we divide the training set into two splits. Concerning the gallery, we evaluate a single template for each identity as the average features vector among all the corresponding face images. Regarding the queries, we randomly select 100 identities, and for each of them, we randomly pick ten correctly classified images, ending up with 1000 queries.

Concerning the learning models, we analyze the performance of two DCNNs: the face classifier from [38] and the CR-trained one from [13]. They share the same structure, i.e., a ResNet-50 [40] architecture equipped with Squeeze-and-Excitation [41] blocks. For both models, we adopt the same preprocessing steps for the images. First, following the same procedure as in Massoli et al. [13], we synthesize different resolution versions of the input that allow us to evaluate the performance of the models in a cross-resolution scenario. Specifically, in our analysis, we consider images at 16, 24, 64, and 256 pixels (shortest side). Next, each image is resized to have the shortest side of 256 pixels, and then it is cropped to a square picture of size 224*x*224 pixels. Finally, we subtract the channel mean from each pixel.

The *SCface* [39] dataset comprises ~ 4K images, shared among 130 different subjects, that have been acquired in an uncontrolled indoor environment. For each person in the dataset, there are five pictures acquired with five different surveillance cameras at three different distances: 1.0, 2.6, and 4.2 m. The three different gaps between the person and the cameras automatically translate into face images with different resolutions. Thus, we use the images contained in the dataset as they are for our purposes. We use the same models and apply the same pre-processing steps as for the VGGFace2 [38] dataset. The only exception is that, in this case, we do not need to down-sampled the images since we already have them available at three different resolutions, each corresponding to a different position of the subject to the camera.

### 4.2. Adversarial attacks

Concerning the generation of the adversarial instances, we exploit the five algorithms we described in Section 3. We use the implementations available in the *foolbox*[1] library with the only exception of the DR one that we build on top of the L-Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [15], optimization procedure. More precisely, the L-BFGS algorithm requires a function to optimize. To our aim, we implement such a function by employing a k-NN algorithm as guidance in the adversarial search. We fit the classifier to the gallery templates we mentioned at the beginning of this section. Then, we start the crafting procedure and stop it as soon as the k-NN classifies the malicious image as belonging to the targeted identity. In Fig. 1, we report a schematic view of the procedure we just described.

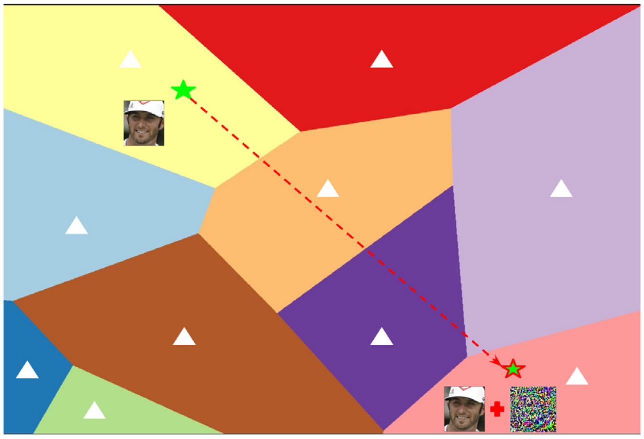

**Fig. 1.** Schematic representation of our approach to crafting DR attacks. The colored regions are the k-NN decision boundaries for ten different identity templates (white triangles). The initial location of the green star represents a correctly classified features vector. The adversarial features vector's final position is represented by the red encircled star. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 4.3. Face identification metrics

FR systems typically deal with sensitive scenarios such as biometric and forensics applications. Hence, different error types have distinct relevance while evaluating system performance, and a simple accuracy measure is not enough to properly evaluate and compare the performance of FR systems. Instead, as mentioned in Section 1, we focus our study on the face identification protocol. Specifically, we consider both the close- and open-set settings.

Concerning the close-set setting, we evaluate the Cumulative Match Characteristic (CMC), a metric that represents a summarized accuracy evaluated on mated searches only, i.e., considering queries that correspond to identities already available the gallery. The CMC value at rank one is usually named “hit rate,” and it is the most typical summary indicator of an algorithm's efficacy. Concerning the VGGFace2 [38] dataset, as we mentioned above, we select 100 identities to construct the queries. Thus, we end up with a gallery containing 8631 identities that comprise a hundred mated ones and 8531 un-mated ones acting as “distractors”.

In the open-set setting, differently from the close-set one, we consider both mated and un-mated queries. To this aim, we remove half of the queries identities from the gallery, ending up with 50 mated and 50 un-mated persons and a gallery containing 8581 templates. With that set, there are two different types of errors that are usually evaluated, i.e., the False Positive Identification Rate (FPIR) and the False Negative Identification Rate (FNIR) or “miss rate”. Concerning the former, it represents the number of un-mated queries that return a positive match at or above a specific similarity threshold. On the other hand, the FNIR represents the number of mated searches that return candidates with a similarity score below the threshold or outside the top R ranks.

The FNIR and FPIR, parametrized by the similarity threshold, can be combined to construct the Detection Error Tradeoff (DET), which is typically used to report the two types of error trade-off. We use the DET to evaluate the performance of the learning models in the experiments.

Finally, concerning the SCface [39] dataset, we evaluate the resilience of the models against attacks at the three different standoff distances. As we will show next in the paper, the results con-

[1] https://foolbox.readthedocs.io/en/stable/

**Table 1**
Attack success rate against classification for “Base” and “Cross-Resolution” models. The first column reports the specific configuration used for each attack. The four values reported in the second and third main columns represent the success rate at a resolution of 16, 24, 64, and 256 pixels, respectively. We emphasize in bold the performance of the strongest attack.

| Attack configuration | Attack success rate (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Base model | | | | Cross-resolution model | | | |
| | 16 | 24 | 64 | 256 | 16 | 24 | 64 | 256 |
| JSMA (1000-0.1) | 76.1 | 61.8 | 25.5 | 11.5 | 65.5 | 62.8 | 17.1 | 6.9 |
| JSMA (1000-0.3) | 96.6 | 92.5 | 75.7 | 61.2 | 96.0 | 94.7 | 70.0 | 50.1 |
| JSMA (1000-0.5) | 98.5 | 95.8 | 86.4 | 76.6 | 97.6 | 97.0 | **100.** | 69.6 |
| CW-$L_2$ (10-10) | 82.9 | 72.9 | 45.9 | 32.7 | 86.4 | 83.3 | 52.8 | 37.4 |
| CW-$L_2$ (10–100) | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** |
| EAD (10-0.1-10) | 95.7 | 98.2 | 94.5 | 87.0 | 96.7 | 99.6 | 98.8 | 98.5 |
| EAD (10-0.1-100) | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** |
| EAD (10-1.0-10) | 83.4 | 85.1 | 50.2 | 27.9 | 72.6 | 94.4 | 86.9 | 73.8 |
| EAD (10-1.0-100) | 98.5 | 99.8 | 98.7 | 91.0 | 97.5 | 99.8 | **100.** | 99.6 |

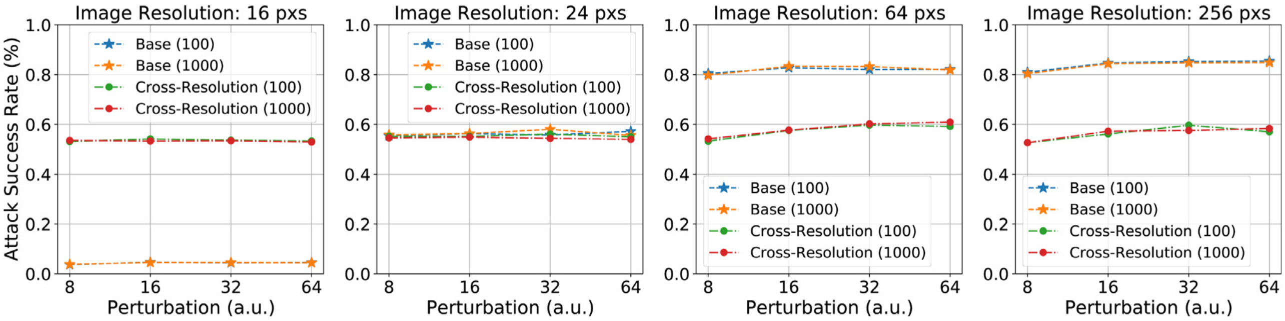

**Fig. 2.** DR [35] attack success rate as function of the maximum allowed perturbation $\delta$ considering 100 and 1000 iteration steps. Each plot represents a different input resolution.

firm the conclusions and intuitions we report about the former dataset.

## 5. Experimental results

We dedicate this section to report the results of our experimental campaigns. As we mentioned in Section 1, we aim to study the behavior of DL-based FR systems when threatened by adversarial attacks in a CR domain. Concerning the FR, as backbone features extractors, we consider the well-known DCNN from Cao et al. [38] that set the state-of-the-art on the NIST datasets [42–44] and the CR model from [13] that set the state-of-the-art in the cross-resolution domain.

To craft adversarial examples, we harness the algorithms we described in Section 3. Moreover, concerning the VGGFace2 [38] dataset, being interested in the CR scenario, we consider input faces at 16, 24, 64, and 256 pixels (shortest side). Concerning the FR task, we keep the gallery at the original resolution. Instead, regarding the SCface [39] dataset, we use the images as they are since they natively represent a cross-resolution domain.

As mentioned in Section 2, to our knowledge, we are the first to conduct this type of study. Thus, a direct comparison with previously published works is not possible. Hence, in what follows, we only report our results. We hope that our study will stimulate further researches in this direction. Throughout this section, we refer to the model from Cao et al. [38] as “Base” model and to the one from Massoli et al. [13] as “Cross-Resolution” model.

In what follows, we first report the results on the VGGFace2 [38] dataset and then on the SCface [39] one.

### 5.1. Threatening the classification

We report the results from the attacks against the classification in Table 1. Concerning the attacks, we use the following configurations. For JSMA, we consider 1000 iterations, a perturbation per pixel equals to 0.1, 0.3, and 0.5 (percentage over the allowed pixel range), and a maximum number of times each pixel can be modified of 10. For CW-$L_2$, we consider 10 binary search steps and 10 and 100 iterations. Concerning EAD, we use the same parameters as for the CW-$L_2$ attack and a value for the weight of the $L_1$ penalty term equals to 0.1 and 1. Furthermore, since the DR [35] attack is the least time demanding compared to the others, we enlarge the set of hyperparameters for it. Thus, we dedicate Fig. 2 to report their results.

From Table 1, we notice that there is no clear signature for which model is more robust against adversarial attacks. On the other hand, we see that, on average, an adversary’s success rate decreases as the resolution increases while keeping the attack configuration fixed. Let us now turn our attention to a single attack, for example, CW-$L_2$. It is interesting to notice the impact of a different choice of hyperparameters. Indeed, even though from the configuration (10-10), the “Base” model seems to be more resilient compared to the “Cross-Resolution” one, this is not true. Indeed, by just increasing the strength of the attack, i.e., (10–100) configuration for which we grow the number of steps, we reach 100% of attack success rate for both models.

From Fig. 2 we observe that it is undeniable that the deep features extracted by the “Cross-Resolution” model are much more robust than those extracted from the “Base” NN. Thus, confirming our previous assertion about the benefit of CR training. From the first plot of Fig. 2, we see that the success rate of the attack is almost 0% for the “Base” model. Instead, in the second plot, it looks like that both models have the same resilience. This is not in contrast with our previous conclusions. Indeed, as it has been shown in appendix 1 of Massoli et al. [13], the “Base” model is not able to generate meaningful deep representation at very low resolutions. Thus, it is almost impossible to craft targeted attacks based on deep features. To sustain even more our assertion, we run a test

**Table 2**
Attacks hit rate. The first column reports the configuration for each attack. The four values reported in the second and third main columns are the results at a resolution of 16, 24, 64, and 256 pixels, respectively As a reference, we report in the first row the hit rate for the authentic images. We emphasize in bold the performance of the strongest attack.

| Attack configuration | Hit rate (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Base model | | | | Cross-resolution model | | | |
| | 16 | 24 | 64 | 256 | 16 | 24 | 64 | 256 |
| Auth | 79.5 | 95.3 | 99.8 | 99.9 | 96.7 | 98.8 | 99.4 | 99.7 |
| JSMA (1000-0.1) | 12.1 | 10.7 | 12.9 | 12.2 | 11.9 | 9.8 | 9.4 | 13.0 |
| JSMA (1000-0.3) | 14.0 | 9.3 | 10.7 | 10.6 | 9.8 | 10.0 | 7.4 | 8.9 |
| JSMA (1000-0.5) | 13.6 | 10.6 | 10.0 | 10.3 | 10.0 | 10.2 | 3.0 | 6.8 |
| CW-$L_2$ (10-10) | 10.9 | 6.5 | 6.1 | 3.7 | 10.8 | 9.3 | 5.5 | 5.1 |
| CW-$L_2$ (10–100) | 7.6 | 4.1 | 6.1 | 2.3 | 9.2 | 9.3 | 3.6 | 4.6 |
| EAD (10-0.1-10) | 31.8 | 32.6 | **27.8** | 25.1 | 19.2 | 16.8 | 19.4 | 19.7 |
| EAD (10-0.1-100) | 17.5 | 9.7 | 6.3 | 6.2 | 13.8 | 11.6 | 6.8 | 5.3 |
| EAD (10-1.0-10) | **44.8** | **38.0** | 26.7 | **25.5** | **20.8** | **25.7** | **20.1** | **21.7** |
| EAD (10-1.0-100) | 34.8 | 30.3 | 20.7 | 16.8 | 17.3 | 16.5 | 17.4 | 17.2 |

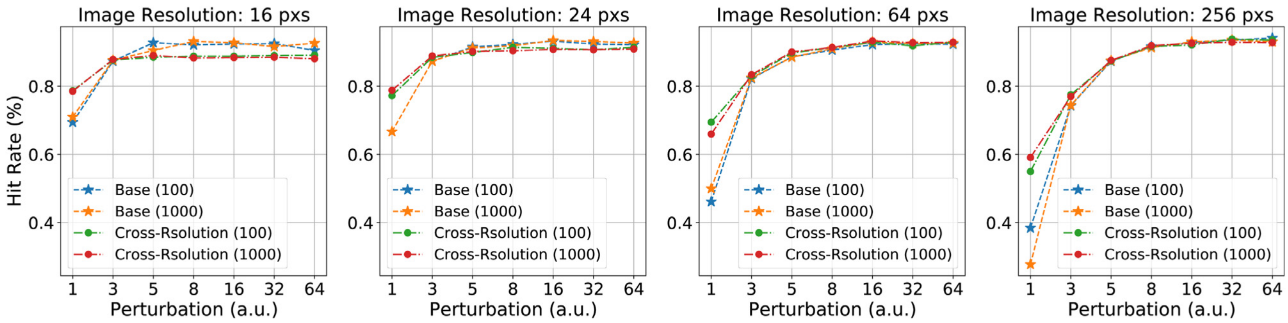

**Fig. 3.** DR [35] hit rate as function of the maximum allowed perturbation $\delta$ considering 100 and 1000 attack steps. Each plot represents a different input resolution.

with untargeted DR attacks in which we easily reach a success rate of 100% for the "Base" model.

Finally, we can notice that from our results, there is no clear evidence in favor of a specific metric since with the proper hyper-parameters, we reached high success rates with the $L_1$, $L_2$, and $L_\infty$.

### 5.2. Threatening the face recognition

We now turn our attention to DL-based FR systems. We begin our analysis by considering the face identification protocol in the close-set scenario, and we then move the open-set one. We refer the reader to Section 4 for a detailed description of the metrics we use to assess the performance of the systems under analysis.

#### 5.2.1. Close-set

As mentioned in Section 4, we use the CMC to evaluate the performance of the threatened models in the close-set scenario. Specifically, we summarize our results in Table 2 by reporting the hit rate, i.e., the CMC value at a rank equals to one, with the exception of the DR [35] attack to which we dedicate Fig. 3. From a defensive point of view, the more resilient a model, the lower the hit rate, while from an attacker perspective, it is the other way round.

By looking at Table 2 and Fig. 3 we can assert that the DR attack is much more effective in fooling a DL-based FR system than the classification-based ones with respect to any type of metric. From the attacker's point of view, this is a fundamental result. Indeed, by comparing the results from Tables 1 and 2, we see that even though the attacks fool the classification, it is not guaranteed that they can evade a similarity-based system. Thus, deep representation attacks might be a better choice to attack an FR system. Moreover, we see how the "Cross-Resolution"-based system exhibits higher robustness than the one based on the "Base" model. Thus, again, we find that DCNNs benefit from a CR training approach [13] in terms of resilience to adversarial attacks. Indeed, it is undeniable that the "Cross-Resolution"-based system is much more resilient against adversarial attacks than the "Base"-based one across all resolutions.

#### 5.2.2. Open-set

To report the results for the face identification protocol in the open-set setting, we exploit the DET. Two fundamental aspects differentiate the DET from the CMC. Indeed, the former applies a threshold among the similarity of the features, and it comprises queries of identities that are not present in the gallery. Instead, the latter does not use any threshold, i.e., it does not discern among "weak" and "strong" similarity scores, and it requires queries related to already known identities.

As we mentioned in Section 4, the DET represents the error trade-off between the FNIR and the FPIR. To summarize the performance of the FR systems, we report the FPIR at a reference value of the FNIR equals to $1.e^{-2}$. Compared to the close-set settings, the adversary's goal is to lower the curve as much as possible, while from a defensive point of view, a higher curve represents a more resilient model. The results are reported in Table 3 with the exception of DR [35] to which we dedicate Fig. 4.

Analyzing the results reported in Table 3 and Fig. 4 we obtain the same conclusions we report for the close-set setting. Specifically, by comparing the results from Table 3 to the ones in Fig. 4 we see that the DR attack is much more effective in fooling the FR system compared to others and that the "Cross-Resolution"-based system is much more resilient than the "Base"-based one against adversarial attacks.

To further confirm our intuition about the dependence of the adversary success from the resolution of the input images and that the robustness of a model against adversarial attacks can benefit from a cross-resolution training procedure, we conduct experiments on the SCface [39] dataset. Specifically, we evalu-

**Table 3**
FPIR@FNIR=1.$e^{-2}$. The first column reports the configuration for each attack. The four values reported in the second and third main columns are the results at a resolution of 16, 24, 64, and 256 pixels, respectively. As a reference, we report in the first row the results for the authentic images. We emphasize in bold the performance of the strongest attack.

| Attack configuration | FPIR@FNIR=1.$e^{-2}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Base model | | | | Cross-resolution model | | | |
| | 16 | 24 | 64 | 256 | 16 | 24 | 64 | 256 |
| Auth | 75.0 | 40.8 | 0.8 | 1.0 | 38.6 | 20.2 | 3.6 | 3.2 |
| JSMA (1000-0.1) | 99.3 | 99.1 | 100. | 95.1 | 99.1 | 98.4 | 100. | 98.1 |
| JSMA (1000-0.3) | 99.0 | 99.1 | 97.2 | 99.7 | 97.8 | 98.6 | 99.0 | 100. |
| JSMA (1000-0.5) | 98.0 | 98.1 | 98.2 | 97.0 | 99.4 | 98.6 | 99.0 | 98.7 |
| CW-$L_2$ (10-10) | 99.5 | 98.1 | 99.5 | 97.4 | 99.0 | **98.1** | 98.9 | 98.9 |
| CW-$L_2$ (10–100) | 100. | 99.0 | 99.5 | 99.4 | 99.6 | **98.1** | 99.6 | 99.2 |
| EAD (10-0.1-10) | **95.3** | **93.2** | 98.7 | 99.5 | 98.4 | 98.8 | **96.0** | 97.6 |
| EAD (10-0.1-100) | 98.0 | 99.4 | 99.4 | 99.0 | 100. | 98.8 | 98.6 | 99.2 |
| EAD (10-1.0-10) | 95.6 | 96.3 | 98.3 | **95.3** | **96.3** | **98.1** | 96.7 | **96.7** |
| EAD (10-1.0-100) | 98.8 | 97.9 | **97.1** | 98.6 | 98.6 | **98.1** | 99.0 | 97.7 |

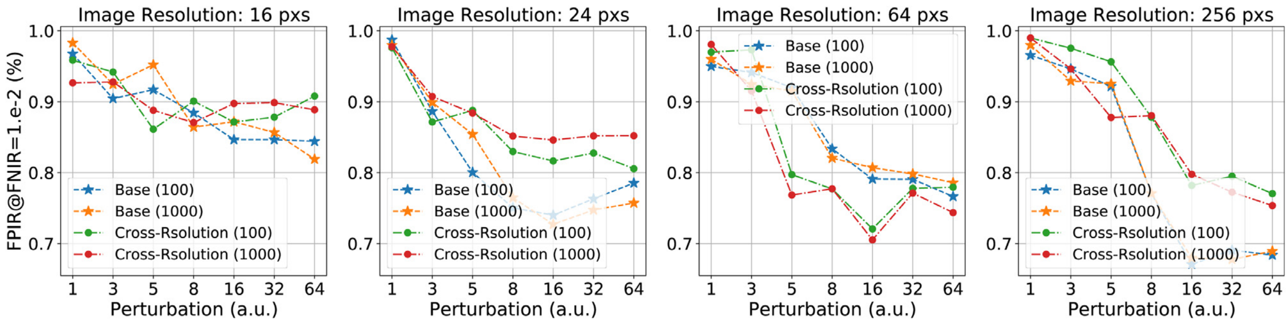

**Fig. 4.** FPIR@FNIR=1.$e^{-2}$ for the DR [35] attack as function of the maximum allowed perturbation $\delta$ considering 100 and 1000 attack steps. Each plot represents a different input resolution.

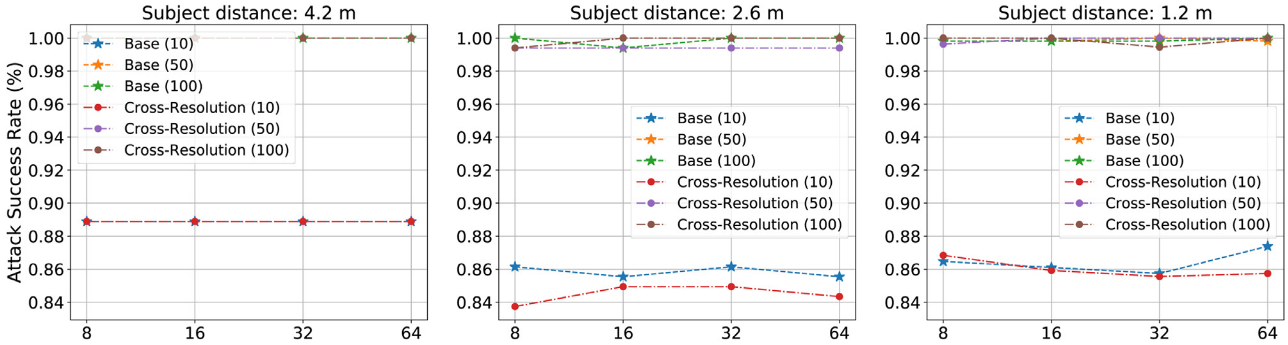

**Fig. 5.** Attack success rate as a function of the maximum allowed perturbation for a different number of iteration steps considering images at the three different standoff distances.

ate the attack success rate for the "Base" model and the "Cross-Resolution" one. Instead of synthesizing low-resolution images, in this case, we consider the images taken at three different distances from the cameras as a down-sampled version of a subject face.

We report the results of the experiments in Table 4 and Fig. 5. In Table 4, the columns d1, d2, and d3 correspond to a distance between the subject and the camera of 4.2, 2.6, and 1.0 m. Also, in this case, we notice a strong correlation between the attack success rate and the input images' resolution. Specifically, we notice that it is easier to fool deep learning models considering low-resolution images than using high-resolution ones. Moreover, we find once again that models can benefit from cross-resolution training.

Finally, in Fig. 5, we report the PGD attack success rate for a different number of iteration steps as a function of the maximum allowed perturbation. We compared the results obtained in the figure, considering 10, 50, and 100 iteration steps. From the figure, we can notice that considering less than 50 iterations, the attack does not converge yet. Instead, above 50 steps, we reach almost 100% of success rate in all cases with the "Cross-Resolution" models still slightly more resilient than the "Base" one.

## 6. Conclusions

DCNN-based FR systems leverage the representation power of learning models. Unfortunately, they also share their weaknesses. Indeed, it has been recently shown that these systems suffer a drastic drop in their performance when tested in a cross-resolution domain. The situation becomes even worse when an adversary comes into play. Indeed, an FR system can be deceived by adversarial examples. These weaknesses pose a severe limit to the spread of these systems to sensitive real-world applications such as biometric systems and forensics.

In such a context, we proposed our analysis in which we compared the resilience to adversarial attacks of FR systems based on the deep features extracted by NNs in a CR scenario. We studied two different DCNN models: a former one, trained only on high-resolution images, and a latter one, trained on a cross-resolution

**Table 4**
Attack success rate against classification for "Base" and "Cross-Resolution" models. The first column reports the specific configuration used for each attack. The columns named d1, d2, and d3 correspond to a distance between the subject and the camera of 4.2, 2.6, and 1.0 m. We emphasize in bold the performance of the strongest attack.

| Attack configuration | Attack success rate (%) | | | | | |
|---|---|---|---|---|---|---|
| | Base model | | | Cross-resolution model | | |
| | d1 | d2 | d3 | d1 | d2 | d3 |
| JSMA (1000-0.1) | 97.2 | 94.6 | 64.7 | 33.3 | 25.9 | 18.5 |
| JSMA (1000-0.3) | **100.** | **100.** | 96.0 | 83.3 | 87.3 | 76.1 |
| JSMA (1000-0.5) | **100.** | **100.** | 99.5 | 94.4 | 95.2 | 90.1 |
| CW-$L_2$ (10-10) | 94.4 | 98.8 | 88.0 | 80.6 | 82.5 | 62.3 |
| CW-$L_2$ (10–100) | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** |
| EAD (10-0.1-10) | **100.** | **100.** | 99.1 | 97.2 | 97.6 | 91.0 |
| EAD (10-0.1-100) | **100.** | **100.** | **100.** | **100.** | **100.** | **100.** |
| EAD (10-1.0-10) | 88.9 | 90.4 | 65.1 | 30.6 | 37.9 | 21.6 |
| EAD (10-1.0-100) | **100.** | **100.** | 97.4 | 83.3 | 91.6 | 80.1 |

domain. To generate adversarial instances, we harnessed several algorithms based on different metrics and objectives, and we craft malicious samples considering input images at a resolution of 16, 24, 64, and 256 pixels for the VGGFace2 dataset and the images at the three different standoff distances concerning the SCface one. Concerning the measures of the performance of the FR systems, we adopted the face identification protocol. Specifically, we considered the close- and open-set settings for which we evaluated the CMC and DET.

From our analysis, concerning the VGGFace2 dataset, we notice that, given a specific configuration, the attack success rate is higher at lower resolutions, for example, at 16 and 24 pixels, than at higher ones, such as 64 and 256 pixels. Such behavior was somehow expected since, at a very low-resolution part of the face, information can be lost, thus simplifying an adversary's effort.

By looking at the FR systems results, it is evident that a DCNN benefits from a CR training procedure since it empowers the learning model to extract more robust deep representations. Moreover, we observed that DR attacks represent a much greater menace to an FR system than the ones based on the classification output of the threatened models for each of the considered metrics, i.e., $L_1$, $L_2$ and $L_\infty$. Such a result was held for the close- as well as for the open-set settings.

Finally, the results we obtain on the SCface dataset confirm all the previous intuitions and conclusions reported following the results obtained on the VGGFace2 dataset.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. This work was partially supported by WAC@Lucca, a project funded by Fondazione Cassa di Risparmio di Lucca (CUP B54I19006080007), and AI4EU, an EC H2020 project (Contract no. 825619).

## References

[1] M. Wang, W. Deng, Deep face recognition: a survey, arXiv:1804.06655(2018).
[2] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: additive angular margin loss for deep face recognition, in: CVPR, IEEE, 2019, pp. 4690–4699.
[3] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans. Image Process. 11 (4) (2002) 467–476.
[4] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE TPAMI (12) (2006) 2037–2041.
[5] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1991, pp. 586–591.
[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[7] F.V. Massoli, G. Amato, F. Falchi, C. Gennaro, C. Vairo, Improving multi-scale face recognition using VGGFace2, in: International Conference on Image Analysis and Processing, Springer, 2019, pp. 21–29.
[8] W.W. Zou, P.C. Yuen, Very low resolution face recognition problem, IEEE Trans. Image Process. 21 (1) (2011) 327–340.
[9] G. Amato, F. Falchi, C. Gennaro, F.V. Massoli, N. Passalis, A. Tefas, A. Trivilini, C. Vairo, Face verification and recognition for digital forensics and information security, in: ISDFS, IEEE, 2019, pp. 1–6.
[10] Z. Cheng, X. Zhu, S. Gong, Surveillance face recognition challenge, arXiv:1804.09691(2018).
[11] H.K. Ekenel, B. Sankur, Multiresolution face recognition, Image Vis. Comput. 23 (5) (2005) 469–477.
[12] X. Luo, Y. Xu, J. Yang, Multi-resolution dictionary learning for face recognition, Pattern Recognit. 93 (2019) 283–292.
[13] F.V. Massoli, G. Amato, F. Falchi, Cross-resolution learning for face recognition, Image Vis. Comput. (2020) 103927.
[14] K. Zhang, Z. Zhang, C.-W. Cheng, W.H. Hsu, Y. Qiao, W. Liu, T. Zhang, Super-identity convolutional neural network for face hallucination, in: European Conference on Computer Vision (ECCV), 2018, pp. 183–198.
[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv:1312.6199(2013).
[16] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: ECML PKDD, Springer, 2013, pp. 387–402.
[17] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, J. Zhu, Efficient decision-based black-box adversarial attacks on face recognition, in: CVPR, IEEE, 2019, pp. 7714–7722.
[18] Q. Song, Y. Wu, L. Yang, Attacks on state-of-the-art face recognition using attentional adversarial attack generative network, arXiv:1811.12026(2018).
[19] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, B. Li, Semanticadv: generating adversarial examples via attribute-conditional image editing, arXiv:1906.07927(2019).
[20] K. Kakizaki, K. Yoshida, Adversarial image translation: Unrestricted adversarial examples in face recognition systems, 2019.
[21] G. Goswami, N. Ratha, A. Agarwal, R. Singh, M. Vatsa, Unravelling robustness of deep learning based face recognition against adversarial attacks, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[22] M. Sharif, S. Bhagavatula, L. Bauer, M.K. Reiter, Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition, in: SIGSAC CCS, ACM, 2016, pp. 1528–1540.
[23] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, arXiv:1607.02533(2016).
[24] K. Sundararajan, D.L. Woodard, Deep learning for biometrics: a survey, ACM Comput. Surv. (CSUR) 51 (3) (2018) 65.
[25] N.A. Spaun, Face recognition in forensic science, in: Handbook of Face Recognition, Springer, 2011, pp. 655–670.
[26] P. Grother, P. Grother, M. Ngan, K. Hanaoka, Face Recognition Vendor Test (FRVT) Part 2: Identification, US Department of Commerce, National Institute of Standards and Technology, 2019.
[27] M. Singh, S. Nagpal, R. Singh, M. Vatsa, A. Majumdar, Magnifyme: aiding cross resolution face recognition via identity aware synthesis, arXiv:1802.08057(2018).
[28] E. Zangeneh, M. Rahmati, Y. Mohsenzadeh, Low resolution face recognition using a two-branch deep convolutional neural network architecture, Expert Syst. Appl. 139 (2020) 112854.
[29] Y. Zhong, W. Deng, Towards transferable adversarial attack against deep face recognition, arXiv:2004.05790(2020).
[30] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, B.Y. Zhao, Fawkes: protecting personal privacy against unauthorized deep learning models, in: Proc. of USENIX Security, 2020.
[31] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: Symposium on Security and Privacy, IEEE, 2017, pp. 39–57.
[32] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, C.-J. Hsieh, Ead: elastic-net attacks to deep neural networks via adversarial examples, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[33] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. 67 (2) (2005) 301–320.
[34] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: 2016 IEEE European Symposium on Security and Privacy, IEEE, 2016, pp. 372–387.
[35] S. Sabour, Y. Cao, F. Faghri, D.J. Fleet, Adversarial manipulation of deep representations, arXiv:1511.05122(2015).
[36] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv:1706.06083(2017).
[37] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv:1412.6572(2014).
[38] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: a dataset for recognising faces across pose and age, in: International Conference on Automatic Face & Gesture Recognition, IEEE, 2018, pp. 67–74.

[39] M. Grgic, K. Delac, S. Grgic, Scface–surveillance cameras face database, Multimed. Tools Appl. 51 (3) (2011) 863–879.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, IEEE, 2016, pp. 770–778.
[41] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks. arXiv, 2017.
[42] B.F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, A.K. Jain, Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, in: CVPR, IEEE, 2015, pp. 1931–1939.
[43] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A.K. Jain, J.A. Duncan, K. Allen, et al., Iarpa janus benchmark-b face dataset, in: CVPR Workshops, IEEE, 2017, pp. 90–98.
[44] B. Maze, J. Adams, J.A. Duncan, N. Kalka, T. Miller, C. Otto, A.K. Jain, W.T. Niggel, J. Anderson, J. Cheney, et al., Iarpa janus benchmark-c: face dataset and protocol, in: ICB, IEEE, 2018, pp. 158–165.