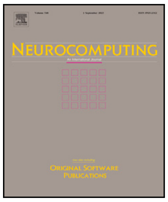
Neurocomputing 580 (2024) 127517

Contents lists available at ScienceDirect

Neurocomputing

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



EAP: An effective black-box impersonation adversarial patch attack method on face recognition in the physical world

Xiaoliang Liu [a,b], Furao Shen [a,c,*], Jian Zhao [d], Changhai Nie [a,b]

[a] *National Key Laboratory for Novel Software Technology, Nanjing University, China*
[b] *Department of Computer Science and Technology, Nanjing University, China*
[c] *School of Artificial Intelligence, Nanjing University, China*
[d] *School of Electronic Science and Engineering, Nanjing University, China*

ARTICLE INFO

Communicated by B. Wu

Dataset link: https://drive.google.com/drive/folders/1w7F2-PnXSXDJQti9x2Bsed5VDrAzmo3t?usp=drive_link

Keywords:
Face recognition
Attacks
Black-box
Adversarial patches
Physical environments

ABSTRACT

Face recognition models and systems based on deep neural networks are vulnerable to adversarial examples. However, existing attacks on face recognition are either impractical or ineffective for black-box impersonation attacks in the physical world. In this paper, we propose EAP, an effective black-box impersonation attack method on face recognition in the physical world. EAP generates adversarial patches that can be printed by mobile and compact printers and attached to the source face to fool face recognition models and systems. To improve the transferability of adversarial patches, our approach incorporates random similarity transformations and image pyramid strategies, increasing input diversity. Furthermore, we introduce a meta-ensemble attack strategy that harnesses multiple pre-trained face models to extract common gradient features. We evaluate the effectiveness of EAP on two face datasets, using 16 state-of-the-art face recognition backbones, 9 heads, and 5 commercial systems. Moreover, we conduct physical experiments to substantiate its practicality. Our results demonstrate that EAP is capable of effectively executing impersonation attacks against state-of-the-art face recognition models and systems in both digital and physical environments.

1. Introduction

Face recognition is a widely used technology that has many applications in our daily lives, such as unlocking phones, verifying payments, and identifying people. However, recent studies have shown that face recognition models based on deep neural networks are vulnerable to adversarial examples, which are maliciously crafted inputs that can fool the models into making wrong predictions. This poses a serious threat to the security and reliability of face recognition systems. Therefore, it is important to study how to generate and defend against adversarial examples for face recognition in the physical world.

Previous works [1–4] on adversarial attacks for face recognition have mainly focused on digital attacks or white-box physical attacks, which require full access to the target model or its gradient information. However, these scenarios are unrealistic in practice, as most face recognition systems are black-boxes that do not reveal their internal details. Moreover, most existing methods [5–7] generate global perturbations on the whole face image, which are hard to implement in the physical world without arousing suspicion. Thus, there is a need for more practical and effective methods that can perform black-box physical impersonation attacks on face recognition using local perturbations.

In this paper, we propose EAP: an Effective Black-Box Impersonation Adversarial Patch Attack Method on Face Recognition in the Physical World. Our method can generate transferable adversarial patches that can be printed by mobile and compact printers and attached to any part of the face to deceive various face recognition models and systems. To achieve this goal, we propose three novel strategies: (1) a random similarity transformation strategy that enhances the diversity of the inputs; (2) an image pyramid strategy that adapts to different scales and resolutions of the inputs; (3) a meta-ensemble attack strategy that extracts common gradient features from multiple pre-trained face models. We evaluate our method on two public face datasets (CelebA-HQ and LFW) using 16 state-of-the-art face recognition backbones and 9 heads. We also test our method on five commercial face recognition systems (Face++, Baidu, Tencent, Microsoft, Huawei). Furthermore, we conduct experiments in a physical environment to verify the effectiveness of our method. The main contributions of our paper are as follows:

- We propose EAP: an effective black-box impersonation attack method on face recognition in the physical world using printed adversarial patches.

[*] Corresponding author.
E-mail addresses: xiaoliang_liu@smail.nju.edu.cn (X. Liu), frshen@nju.edu.cn (F. Shen), jianzhao@nju.edu.cn (J. Zhao), changhainie@nju.edu.cn (C. Nie).

https://doi.org/10.1016/j.neucom.2024.127517
Received 11 June 2023; Received in revised form 22 January 2024; Accepted 4 March 2024
Available online 6 March 2024
0925-2312/© 2024 Elsevier B.V. All rights reserved.

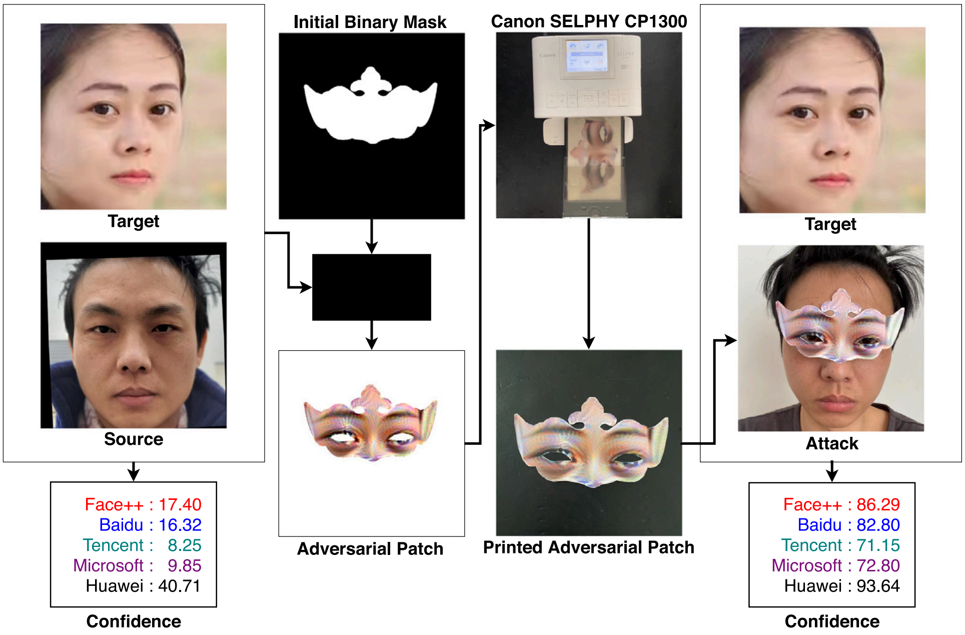


Fig. 1. An example of EAP attack on the commercial face recognition systems in the physical world.

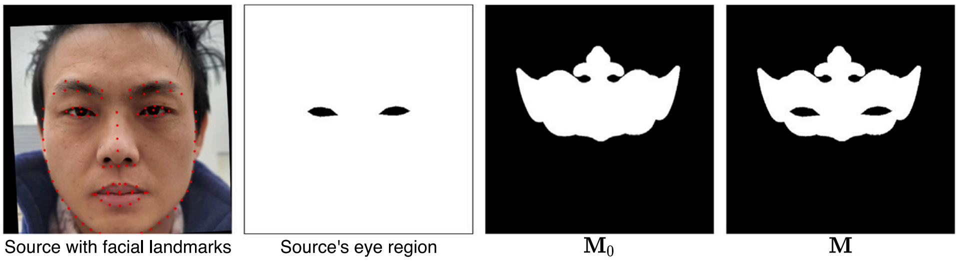


Fig. 2. An example of generating a binary mask $\boldsymbol{M}$ using the initial binary mask $\boldsymbol{M}_0$ and the facial landmarks of the source.

- We propose three novel strategies (random similarity transformation strategy; image pyramid strategy; meta-ensemble attack strategy) that improve the transferability of adversarial patches.
- We perform extensive experiments on public datasets and commercial systems using various settings and metrics. Experiments demonstrate that EAP is an effective black-box impersonation attack method on face recognition in the physical world. Particularly, EAP is able to perform effective impersonation attacks on commercial face recognition systems using a mobile and compact printer. Fig. 1 shows an example of EAP attack on the commercial face recognition systems in the physical world.

2. Related works

In this section, we review the existing literature on adversarial attacks on face recognition, especially those that can be implemented in the physical world. We also discuss some methods that aim to improve the transferability of adversarial examples across different models and systems. We highlight the main challenges and limitations of the current approaches and how our proposed method addresses them.

Adversarial attacks on face recognition can be classified into two categories: digital attacks and physical attacks. Digital attacks generate perturbed images that can fool face recognition models in a digital environment, such as a computer screen or a camera. Physical attacks generate perturbed objects or accessories that can fool face recognition models in a real-world environment, such as a printed photo or a pair of glasses.

Digital attacks can be further divided into two types: white-box attacks and black-box attacks. White-box attacks assume full access to the target model, such as its architecture and parameters, and use gradient-based methods to generate adversarial examples (Goodfellow et al. [8]; Madry et al. [9]; Carlini and Wagner [10]). Black-box attacks assume limited or no access to the target model, and use query-based methods (Chen et al. [11]; Ilyas et al. [12]), transfer-based methods (Liu et al. [13]; Dong et al. [14]), or generative methods (Song et al. [15]; Yang et al. [7]) to generate adversarial examples.

Physical attacks are more challenging than digital attacks because they need to consider various factors that affect the appearance of the perturbed object in different scenarios, such as lighting conditions, viewing angles, distances, etc. Some physical attack methods use eyeglasses (Sharif [1,2]), hats (Komkov and Petiushko [4]), masks (Zolfi et al. [16], Yin et al. [17]), stickers (Wei et al. [18]) or light projection (Nguyen et al. [3]) as carriers of adversarial perturbations. However, these methods either require full access to the target model or its gradient information, or rely on specific equipment that may not be available or convenient in practice.

Another challenge for physical attack methods is to achieve high transferability across different face recognition models and systems.

Transferability refers to the property that an adversarial example generated for one model can also fool another model. Transferability is important for black-box physical attack methods because they usually do not have access to the target model or system. Some methods try to improve transferability by using ensemble-based strategies (Liu et al. [13]; Dong et al. [14]; Xiao et al. [19]), which combine multiple pre-trained models to generate more generalizable adversarial examples. However, these methods still have limitations in terms of impersonation performance and applicability to commercial face recognition systems.

In this paper, we propose a novel black-box physical impersonation attack method on face recognition, called EAP. EAP uses printed patches as carriers of adversarial perturbations that can be easily attached to any part of the face. EAP also uses random similarity transformation and image pyramid strategies to enhance the diversity and robustness of the inputs for generating adversarial patches. Moreover, EAP uses a meta-ensemble attack strategy that leverages ensemble-learning techniques to extract more common gradient features from multiple pre-trained models for generating more transferable adversarial patches. EAP can effectively perform impersonation attacks on various state-of-the-art face recognition models and systems in both digital and physical environments.

3. Methodology

In this section, we provide the detailed description of our algorithm. We first introduce the adversarial patch attack formulation. To increase the diversity of the inputs and improve the transferability of the adversarial patches, we then introduce a random similarity transformation strategy and an image pyramid strategy. Finally, we introduce a meta-ensemble attack strategy that leverages several pre-trained face models to extract more common gradient features.

3.1. Problem formulation

An adversarial patch is a manipulated small region in an image, which can deceive a face recognition model. Attackers can wear such patches on clothing items, stickers, or masks to impersonate other people or evade detection. In order to define the problem, we represent the face image of a target identity with $\boldsymbol{x}^{(t)}$, the attacker's source image with $\boldsymbol{x}^{(s)}$ and the attack image that contains an adversarial patch with $\bar{\boldsymbol{x}}$. The objective of the adversarial patch attack is to find a solution for the following optimization problem with a constraint:

$$\begin{aligned}&\text{argmin}_{\bar{\boldsymbol{x}}} \mathcal{L}(f(\bar{\boldsymbol{x}}), f(\boldsymbol{x}^{(t)})),\\&\text{s.t. } \bar{\boldsymbol{x}} \odot (1-\boldsymbol{M}) = \boldsymbol{x}^{(s)} \odot (1-\boldsymbol{M}),\end{aligned} \tag{1}$$

where $f(\boldsymbol{x}) : \boldsymbol{X} \to \mathbb{R}^d$ is a face recognition model that extracts a normalized feature representation vector for an input image $\boldsymbol{x} \in \boldsymbol{X}$, $\odot$ is the element-wise product operator, and $\boldsymbol{M}$ is a binary mask. The binary mask $\boldsymbol{M}$ serves to limit pixel perturbations to cases where the corresponding position in $\boldsymbol{M}$ is set to 1. Fig. 2 displays an example of the binary mask $\boldsymbol{M}$ generated using the initial binary mask $\boldsymbol{M}_0$ and the facial landmarks of the source image. We detect the facial landmarks using PIPNet [20]. The loss function $\mathcal{L}$ is a cosine similarity loss function given by,

$$\mathcal{L}(\boldsymbol{v}_1, \boldsymbol{v}_2) = 1 - cos(\boldsymbol{v}_1, \boldsymbol{v}_2). \tag{2}$$

where $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ denote the feature vectors extracted by the face recognition feature extractor.

3.2. Random similarity transformation strategy

To enhance the diversity of the inputs and thus the transferability of adversarial patches, we propose a random similarity transformation (RST) strategy that applies random scaling, rotation, and translation to each input image before feeding it into the face recognition model. The RST strategy is controlled by a single hyperparameter, denoted as β, which determines the range of the random similarity transformation with four degrees of freedom (4DoF). The transformation parameters are sampled uniformly from the following intervals:

$$\begin{aligned}t_x &= U(-\beta W, \beta W),\\t_y &= U(-\beta H, \beta H),\\\theta &= U(-\beta\pi/2, \beta\pi/2),\\s &= U(1-\beta, 1+\beta),\end{aligned} \tag{3}$$

where W and H are the width and height of the input image, and $U(a,b)$ denotes a uniform distribution over the interval $[a,b]$. Using these parameters, we construct a random similarity transformation matrix $\boldsymbol{T}$ as follows:

$$\begin{aligned}\boldsymbol{T} &= \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} cos(\theta) & sin(\theta) & 0 \\ -sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}\\&= \begin{bmatrix} s \cdot cos(\theta) & s \cdot sin(\theta) & t_x \\ -s \cdot sin(\theta) & s \cdot cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix},\end{aligned} \tag{4}$$

Given a coordinate (p_x, p_y) of the input image, we can obtain its transformed coordinate (p'_x, p'_y) by applying the matrix $\boldsymbol{T}$:

$$\begin{bmatrix} p'_x \\ p'_y \\ 1 \end{bmatrix} = \boldsymbol{T}\begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix}. \tag{5}$$

Finally, we generate the transformed input image by bilinear interpolation.

3.3. Image pyramid strategy

To further enhance the transferability of our adversarial patches, we propose an image pyramid (IP) strategy that can adapt to different scales and resolutions of the input images of the face recognition model and system. We denote the γ levels image pyramid as $\bar{\boldsymbol{X}} = \{\bar{\boldsymbol{x}}^{(1)}, \bar{\boldsymbol{x}}^{(2)}, \ldots, \bar{\boldsymbol{x}}^{(\gamma)}\}$, where each level i has a height H_i and a width W_i. These are defined as:

$$H_i = (H_0 + \frac{H_0}{2} \times (i-1)), W_i = (W_0 + \frac{W_0}{2} \times (i-1)), \tag{6}$$

where H_0 and W_0 are the height and width of the target model input.

Our RST and IP strategies combined simulate various poses, distances, camera angles, and focal lengths of faces in real scenes. This allows our method to learn more general features in the pre-trained model, improving the transferability of the adversarial patches. The detailed algorithm of EAP is described in Algorithm block 1, and an example of one iteration of EAP is shown in Fig. 3.

3.4. Meta-ensemble attack strategy

In this subsection, we introduce a meta-ensemble attack strategy that can improve the transferability of adversarial patches by ensembling several pre-trained face models. The idea is to extract more common gradient features from different face models and use them to generate more robust adversarial patches that can fool multiple face models and systems.

Ensemble methods have been widely applied in research and competitions to improve performance and robustness [21,23,24]. However, using a hard ensemble method, as proposed by Dong et al. can overfit the pre-trained face models, as the number of available models is limited. To address this issue, we propose a meta-ensemble attack strategy, inspired by the meta-learning framework for few-shot learning [25–27].

As illustrated in Fig. 4, compared to the one-stage hard-ensemble attack, the meta-ensemble attack involves two stages of updates. Let $F = \{f_1, f_2, \ldots, f_m\}$ denote the m target pre-trained face models. First, we need to randomly select k models from F,

$$\{p_1, p_2, \ldots, p_k\} = \text{Random.choice}(F, k), 1 \le k < m. \tag{7}$$

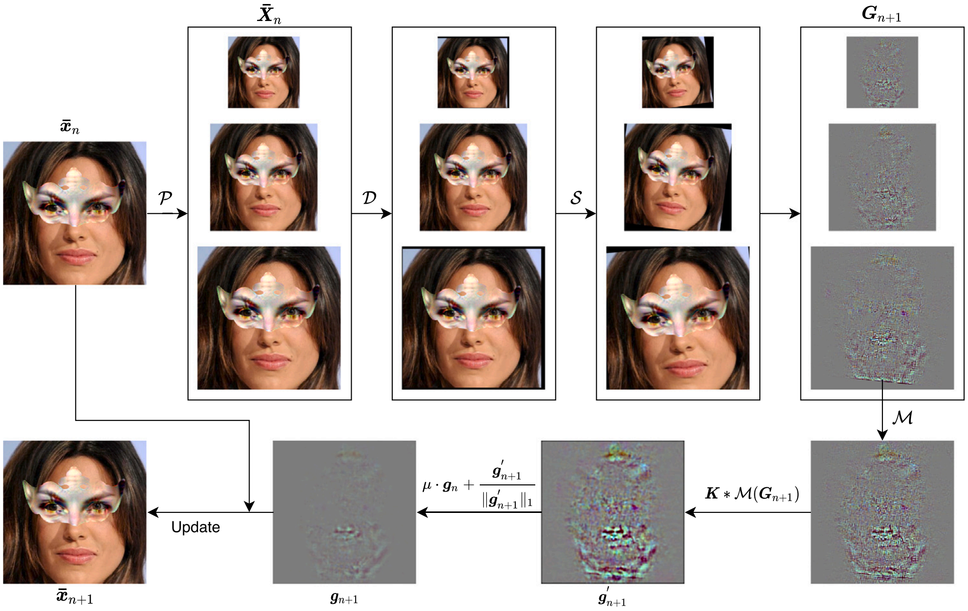


Fig. 3. An example of one iteration of EAP. $\mathcal{P}$ denotes the image pyramid function. $\mathcal{D}$ denotes the diversity transformation function. $\mathcal{S}$ denotes the random similarity transformation function. $\boldsymbol{G}_{n+1}$ are the gradients of $\bar{\boldsymbol{X}}_n$. $\mathcal{M}$ is the mean function. $\boldsymbol{K}$ is the Gaussian kernel. μ is the decay factor.

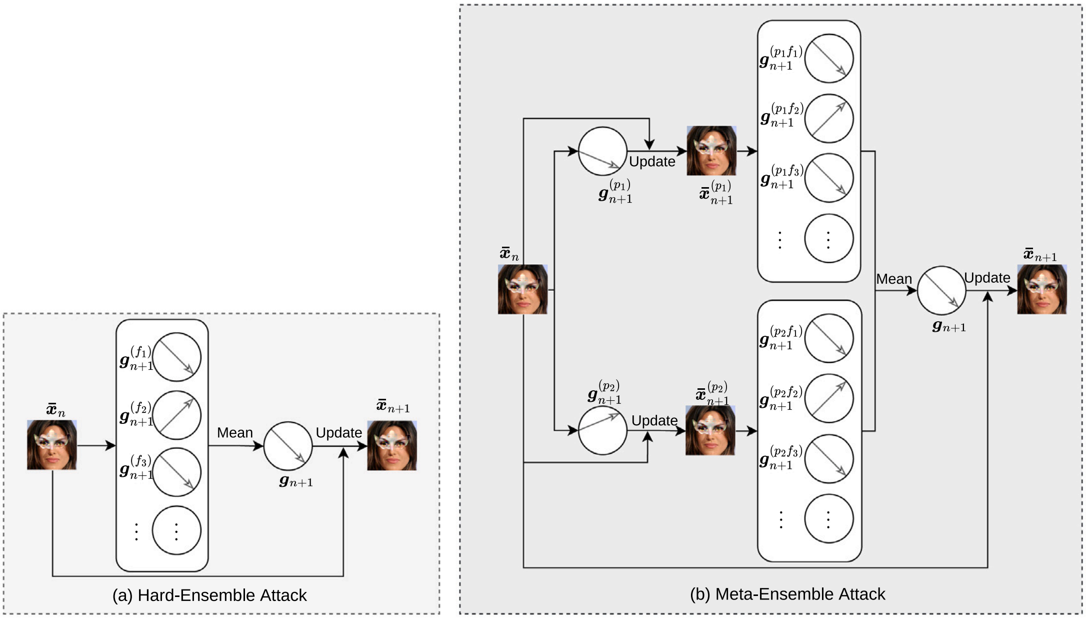

$\bar{\boldsymbol{x}}_n$
$\boldsymbol{g}^{(f_1)}_{n+1}$
$\boldsymbol{g}^{(f_2)}_{n+1}$
$\boldsymbol{g}^{(f_3)}_{n+1}$
Mean
Update
$\boldsymbol{g}_{n+1}$
$\bar{\boldsymbol{x}}_{n+1}$
(a) Hard-Ensemble Attack
$\bar{\boldsymbol{x}}_n$
$\boldsymbol{g}^{(p_1)}_{n+1}$
Update
$\bar{\boldsymbol{x}}^{(p_1)}_{n+1}$
$\boldsymbol{g}^{(p_1 f_1)}_{n+1}$
$\boldsymbol{g}^{(p_1 f_2)}_{n+1}$
$\boldsymbol{g}^{(p_1 f_3)}_{n+1}$
$\boldsymbol{g}^{(p_2)}_{n+1}$
Update
$\bar{\boldsymbol{x}}^{(p_2)}_{n+1}$
$\boldsymbol{g}^{(p_2 f_1)}_{n+1}$
$\boldsymbol{g}^{(p_2 f_2)}_{n+1}$
$\boldsymbol{g}^{(p_2 f_3)}_{n+1}$
Mean
Update
$\boldsymbol{g}_{n+1}$
$\bar{\boldsymbol{x}}_{n+1}$
(b) Meta-Ensemble Attack

Fig. 4. An example of one iteration of ensemble attacks. (a) The hard-ensemble attack requires one phase of update. (b) The meta-ensemble attack requires two phases of updates.

Then, we use the k models to obtain k gradients $\{\boldsymbol{g}^{(p_1)}_{n+1}, \boldsymbol{g}^{(p_2)}_{n+1}, \ldots, \boldsymbol{g}^{(p_k)}_{n+1}\}$ and use each of these k gradients to update $\bar{\boldsymbol{x}}_n$ in the first stage to obtain $\{\bar{\boldsymbol{x}}^{(p_1)}_{n+1}, \bar{\boldsymbol{x}}^{(p_2)}_{n+1}, \ldots, \bar{\boldsymbol{x}}^{(p_k)}_{n+1}\}$. Next, we use $\{\bar{\boldsymbol{x}}^{(p_1)}_{n+1}, \bar{\boldsymbol{x}}^{(p_2)}_{n+1}, \ldots, \bar{\boldsymbol{x}}^{(p_k)}_{n+1}\}$ and F to obtain $k \times m$ gradients and average the $k \times m$ gradients to obtain $\boldsymbol{g}_{n+1}$. Finally, we use $\boldsymbol{g}_{n+1}$ to update $\bar{\boldsymbol{x}}_n$ in the second stage to obtain $\bar{\boldsymbol{x}}_{n+1}$.

Compared to the one-stage hard-ensemble attack, the two-stage meta-ensemble attack makes the number of gradients improve from m to $k \times m$. Therefore, the meta-ensemble attack strategy improves the transferability of adversarial patches by leveraging the common features of multiple pre-trained face models.

4. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of EAP on face recognition models and systems. We first introduce the datasets, models and metrics used in our experiments. Then, we compare EAP with existing methods on digital attacks. Next, we analyze the impact of different components and strategies of EAP on the attack performance. Finally, we demonstrate the feasibility and robustness of EAP on commercial face recognition systems in the physical world.

4.1. Experimental setup

Datasets. We use two public face datasets for the experiments: CelebA-HQ [28], which has high-quality faces, and LFW [29], which has low-resolution faces. We randomly choose 500 pairs of different identities from each dataset to measure the attack performance.

Face Recognition Backbones, Heads and Systems. In our experiments, we mainly use state-of-the-art 16 backbones, 9 face recognition

Algorithm 1 EAP algorithm

```
Require: A face image $\boldsymbol{x}^{(t)}$ of the target identity; a source image $\boldsymbol{x}^{(s)}$ of
    the attacker; the initial binary mask $\boldsymbol{M}_0$; the facial landmarks $lms^{(s)}$
    of the source image; a target face model $f$;
Require: The number of iterations $N$; the perturbation bound $\epsilon$; the
    perturbation step size $\alpha$; the hyperparameter $\beta$ for the random sim-
    ilarity transformation; the number of levels of the image pyramid
    $\gamma$; the decay factor $\mu$ as in [21].
Require: The binary mask generation function $\mathcal{G}$; the image pyramid
    function $\mathcal{P}$; the diversity transformation function $\mathcal{D}$ as in [22]; the
    random similarity transformation function $\mathcal{S}$; the mean function $\mathcal{M}$;
    the sign function $\mathrm{sign}$.
Require: The cosine similarity loss function $\mathcal{L}$.
$\boldsymbol{M} \leftarrow \mathcal{G}(\boldsymbol{M}_0, lms^{(s)})$;
$\bar{\boldsymbol{x}}_0 \leftarrow \boldsymbol{x}^{(s)} \odot (1 - \boldsymbol{M}) + \boldsymbol{x}^{(t)} \odot \boldsymbol{M}$;
$\boldsymbol{g}_0 \leftarrow 0$;
for $n \leftarrow 0$ to $N - 1$ do
    Obtain image pyramid,
        $\bar{\boldsymbol{X}}_n = \{\bar{\boldsymbol{x}}_n^{(1)}, \bar{\boldsymbol{x}}_n^{(2)}, \cdots, \bar{\boldsymbol{x}}_n^{(\gamma)}\} \leftarrow \mathcal{P}(\bar{\boldsymbol{x}}_n, \gamma)$;
    Obtain the gradients of the image pyramid,
        $\boldsymbol{G}_{n+1} = \{\boldsymbol{g}_{n+1}^{(1)}, \boldsymbol{g}_{n+1}^{(2)}, \cdots, \boldsymbol{g}_{n+1}^{(\gamma)}\} \leftarrow \nabla_{\bar{X}_n} \mathcal{L}(f(\mathcal{S}(\mathcal{D}(\bar{\boldsymbol{X}}_n), \beta)), f(\mathbf{x}^{(t)}))$;
    Update gradient,
        $\boldsymbol{g}_{n+1} \leftarrow \mu \cdot \boldsymbol{g}_n + \frac{\boldsymbol{K} * \mathcal{M}(\boldsymbol{G}_{n+1})}{\|\boldsymbol{K} * \mathcal{M}(\boldsymbol{G}_{n+1})\|_1}$,
    where $\boldsymbol{K}$ is the Gaussian kernel and $*$ is convolution;
    Update $\bar{\boldsymbol{x}}_{n+1}$ by applying the sign gradient as
        $\bar{\boldsymbol{x}}_{n+1} \leftarrow \underset{[\bar{\boldsymbol{x}}_0 - \epsilon, \bar{\boldsymbol{x}}_0 + \epsilon]}{\mathrm{Clip}} (\bar{\boldsymbol{x}}_n - \alpha \cdot \mathrm{sign}(\boldsymbol{g}_{n+1}) \odot \boldsymbol{M})$;
end for
return $\bar{\boldsymbol{x}}_N$
```

heads, and 5 commercial face recognition systems to evaluate the performance of the attacks.

Backbones: Swin-T [30], EfficientNet-B0 (Effi-B0) [31], Resnet50 (IR50) [32], Resnet152-SE (IRSE151) [33], TFNAS-A [34], ReXNetV1 [35], ResNeSt50 [36], RepVGG-A0 [37], RepVGG-B0, RepVGG-B1, MobileFaceNet (MFNet) [38], LightCNN29 [39], HRNet [40], GhostNet [41], Attention56 (Att56) [42], and Attention92 (Att92).

Heads: AdaCos (AdaC) [43], AdaM-Softmax (AdaMS) [44], AM-Softmax (AMS) [45], ArcFace (ArcF) [46], CircleLoss (CircleL) [47], CurricularFace (CurriF) [48], MagFace (MagF) [49], MV-Softmax (MVS) [50], and NPCFace (NPCF) [51].

Systems: Face++ [52], Baidu [53], Tencent [54], Microsoft [55], and Huawei [56]. In Microsoft, we use the “recognition_04” face recognition model and the “detection_03” face detection model, which are the latest versions of Microsoft’s face recognition system. All other face recognition systems use the default version.

The pre-trained models used for evaluation in the experiments come from FaceX-Zoo [57].

Competitors. We chose the global adversarial example attack methods, *e.g.*, PGD [9], DIM [22] and TIDIM [14], and the adversarial patch attack methods, *e.g.*, PGDAP [58] and TAP [19], as the main comparison methods.

Evaluate Metrics. For impersonation attacks on the face recognition models, the attack success rate (ASR) [6,17,59] is reported as an evaluation metric,

$$ASR = \frac{\sum_{i=1}^{N} 1_{\tau}(cos[f(\boldsymbol{x}_i^{(t)}), f(\bar{\boldsymbol{x}}_i)] > \tau)}{N} \times 100\%, \tag{8}$$

where N is the number of pairs in the face dataset, 1_τ denotes the indicator function, τ is a pre-determined threshold. For each victim facial recognition model, τ will be determined at 0.001 False Acceptance Rate (FAR) on all possible image pairs in LFW. $\boldsymbol{x}^{(t)}$ is the target and $\bar{\boldsymbol{x}}$ is the attack.

For the evaluation of the attacks on the face recognition systems, we report the mean confidence scores (MCS) on each dataset as an evaluation metric,

$$MCS = \frac{\sum_{i=1}^{N} conf_i}{N} \times 100\%, \tag{9}$$

where *conf* is a confidence score between the target and the attack returned from the face recognition system API, N is the number of pairs in the face dataset.

Additionally, we implement our codes based on the open source deep learning platform PyTorch [60].

4.2. Digital-environment experiments

In this subsection, we present the outcomes of the black-box impersonation attacks in the digital world.

Attacking the Backbones. We first evaluate the effectiveness of EAP on attacking the face recognition backbones. We conducted an experiment using Swin-S [30] as the target face model f, and MV-Softmax as the face recognition head for all backbones. We set the number of iterations N for all attack methods to 1000. For global adversarial example attack methods, we set the perturbation step size α to 0.007 and the perturbation bound ϵ to 0.031. For adversarial patch attack methods, we set the perturbation step size α to 0.003 and the perturbation bound ϵ to 0.3. In addition, we set the decay factor μ to 0.4 for DIM, TIDIM, TAP, and EAP. We also set the hyperparameter β to 0.1 and the number of levels of the image pyramid γ to 3 for EAP.

The results are reported in Table 1, which demonstrates that EAP outperforms TAP, PGDAP, PGD, DIM, and TIDIM for impersonation attacks on both CelebA-HQ and LFW datasets. Specifically, EAP achieves an average ASR of 91.22% on CelebA-HQ and 90.76% on LFW, while TAP only achieves 60.46% and 62.64% for black-box impersonation attacks. Additionally, we observe that EAP can achieve high attack success rates on all backbones, ranging from 82.6% to nearly 100%. These results indicate that EAP can effectively perform black-box impersonation attacks on various face recognition backbones with different architectures and complexities, including transformer-based models, efficient models, and residual models.

Attacking the Heads. We also evaluate the effectiveness of EAP on attacking the face recognition heads. The results are shown in Table 2. The backbones are set as MobileFaceNet. The other settings are the same as in the attacking the backbones.

As shown in Table 2, EAP outperforms other attack methods on all victim heads, regardless of the quality or resolution of the face images in CelebA-HQ or LFW datasets. This demonstrates the effectiveness of EAP on different head architectures.

Attacking the Commercial Systems. To evaluate the effectiveness of EAP on commercial face recognition systems, we conducted experiments on five popular online services: Face++, Baidu, Tencent, Microsoft and Huawei. The results are shown in Table 3. EAP-H is the EAP with the hard-ensemble attack. EAP-M is the EAP with the meta-ensemble attack. For the EAP-H and EAP-M ensemble attacks, the target pre-trained face models F of the ensemble attacks use the Swin-S, Swin-T, Effi-B0, IR50, IRSE151, TFNAS-A, ReXNetV1, ResNeSt50, RepVGG-A0, RepVGG-B0, RepVGG-B1, MFNet, LightCNN29, HRNet, GhostNet, Att56, and Att92 with MV-Softmax and Resnet101-SE, Resnet50-SE, Resnet151, and MFNet with ArcFace. The number of iterations N in the EAP-H and EAP-M is set to 200. The hyperparameter k in the EAP-M is set to 2. The other settings are the same as in the attacking the backbones.

Table 1
The results of the black-box impersonation attacks on the SOTA backbones.

Backbone	CelebA-HQ						LFW					
	Global			Patch			Global			Patch		
	PGD	DIM	TIDIM	PGDAP	TAP	EAP	PGD	DIM	TIDIM	PGDAP	TAP	EAP
Swin-T	77.00	97.40	95.20	47.00	95.40	**99.80**	82.40	99.40	**99.80**	13.20	90.60	99.60
Effi-B0	4.40	32.20	70.40	31.00	54.60	**87.20**	0.20	24.80	75.20	9.40	54.40	**85.60**
IR50	2.00	37.40	79.00	24.60	56.40	**90.40**	0.00	32.00	89.20	7.20	61.40	**91.40**
IRSE152	3.80	42.60	82.80	26.40	58.00	**93.40**	0.60	43.60	92.80	7.80	63.60	**93.60**
TFNAS-A	4.00	45.40	81.80	38.20	54.80	**94.00**	0.00	35.40	84.80	14.40	56.60	**92.40**
ReXNetV1	5.00	38.60	78.20	36.60	64.00	**94.20**	0.40	31.20	84.80	14.60	66.00	**90.60**
ResNeSt50	3.20	41.00	82.40	26.80	41.40	**88.20**	0.20	33.00	**90.00**	5.60	25.40	83.40
RepVGG-A0	4.60	37.00	71.20	34.00	57.60	**86.80**	0.40	36.20	86.20	18.40	72.40	**90.40**
RepVGG-B0	2.20	32.40	72.60	28.40	62.60	**85.20**	0.40	39.60	89.20	13.80	76.80	**93.40**
RepVGG-B1	3.60	38.40	77.40	26.60	61.00	**88.40**	0.60	40.80	90.40	12.00	75.00	**94.60**
MFNet	7.00	34.80	71.00	38.00	53.20	**86.40**	0.40	23.20	75.20	17.60	54.00	**83.20**
LightCNN29	3.20	25.40	60.20	40.20	60.40	**86.80**	0.00	16.80	64.80	19.80	58.20	**82.60**
HRNet	3.00	40.20	82.80	25.80	54.20	**95.00**	0.60	35.40	87.20	10.60	53.20	**92.60**
GhostNet	4.00	33.20	67.40	35.80	51.60	**88.60**	0.00	23.60	73.60	18.00	48.40	**85.20**
Att56	4.00	52.20	90.40	37.60	74.80	**98.20**	1.00	52.60	96.80	15.60	75.60	**97.40**
Att92	2.80	45.60	86.20	25.60	67.40	**97.00**	0.40	47.60	94.20	10.00	70.60	**96.20**
Mean	8.36	42.11	78.06	32.66	60.46	**91.22**	5.48	38.45	85.89	13.00	62.64	**90.76**

Table 2
The results of the black-box impersonation attacks on the SOTA heads.

Dataset: CelebA-HQ											
Method		AdaC	AdaMS	AMS	ArcF	CircleL	CurriF	MagF	MVS	NPCF	Mean
Global	PGD	5.00	6.40	3.80	3.40	11.20	5.00	4.40	7.00	3.00	5.47
	DIM	21.80	34.60	27.60	23.40	45.40	28.20	35.20	34.80	20.20	30.13
	TIDIM	54.80	68.60	60.40	60.80	75.80	61.20	69.80	71.00	56.00	64.27
Patch	PGDAP	32.20	38.80	32.40	26.60	55.40	34.40	37.60	38.00	27.40	35.87
	TAP	49.20	62.00	46.80	45.80	65.20	55.60	55.20	53.20	41.00	52.67
	EAP	**72.60**	**88.00**	**79.80**	**75.60**	**88.80**	**82.00**	**82.60**	**86.40**	**72.80**	**80.96**
Dataset: LFW											
Global	PGD	0.00	0.20	0.00	0.00	0.60	0.20	0.40	0.40	0.00	0.20
	DIM	10.20	22.00	14.80	14.60	31.60	16.40	26.00	23.20	12.60	19.04
	TIDIM	51.80	72.60	65.00	65.40	75.80	63.40	73.40	75.20	63.20	67.31
Patch	PGDAP	8.60	17.00	9.20	10.80	30.00	14.00	16.40	17.60	8.80	14.71
	TAP	36.80	56.00	39.60	44.00	59.20	49.20	51.40	54.00	37.60	47.53
	EAP	**62.00**	**85.00**	**73.20**	**72.20**	**86.20**	**77.40**	**81.00**	**83.20**	**68.80**	**76.56**

Table 3
The results of the black-box impersonation attacks on the SOTA commercial face recognition systems.

Dataset: CelebA-HQ							
Method		Face++	Baidu	Tencent	Microsoft	Huawei	Mean
Global	PGD	45.18	31.06	15.46	12.09	50.73	30.91
	DIM	57.37	47.69	28.45	28.98	72.28	46.95
	TIDIM	66.32	58.08	40.05	42.09	83.63	58.03
Patch	PGDAP	61.77	53.07	24.51	20.77	64.63	44.95
	TAP	65.04	54.97	22.17	20.35	66.49	45.80
	EAP	70.25	65.77	34.58	37.84	78.96	57.48
	EAP-H	75.76	73.75	50.47	54.47	90.11	68.91
	EAP-M	**75.94**	**73.89**	**50.69**	**54.67**	**90.23**	**69.08**
Dataset: LFW							
Global	PGD	38.03	21.87	15.25	10.96	48.33	26.89
	DIM	52.56	43.30	28.38	31.72	72.29	45.65
	TIDIM	64.78	61.61	43.58	53.79	84.95	61.74
Patch	PGDAP	54.64	41.15	20.70	14.83	55.52	37.37
	TAP	60.18	50.86	19.56	16.50	61.69	41.76
	EAP	66.12	62.90	32.68	36.21	76.87	54.96
	EAP-H	71.88	71.97	50.07	54.28	**89.54**	67.55
	EAP-M	**71.92**	**72.12**	**51.14**	**54.46**	89.49	**67.83**

As shown in Table 3, we can see that EAP-M can effectively attack all tested commercial systems with high MCS, ranging from 50.69% to 90.23%. Specifically, EAP-M achieves the highest MCS of 75.94% and 71.92% on the Face++, which is a widely used commercial system. Furthermore, the visualization results are shown in Fig. 5. We can see from Fig. 5 that the confidence scores between the targets and the attacks generated via EAP are much higher than the others. Moreover, EAP-M achieved confidence scores of over 70% on all commercial systems. These results indicate that EAP can pose a serious threat to the security of commercial face recognition systems.

We also observed that some services are more robust than others against EAP attacks. For example, Tencent and Microsoft has relatively

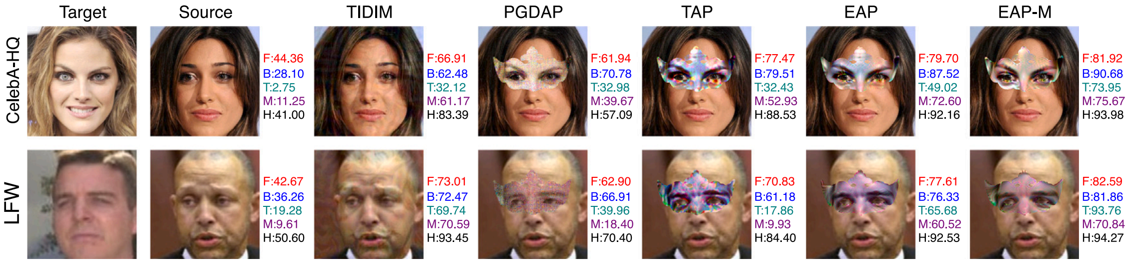


Fig. 5. The visualization results of digital black-box impersonation attacks on five commercial face recognition systems. The confidence scores are pasted to the right of each attack, (F:Face++, B:Baidu, T:Tencent, M:Microsoft, H:Huawei).

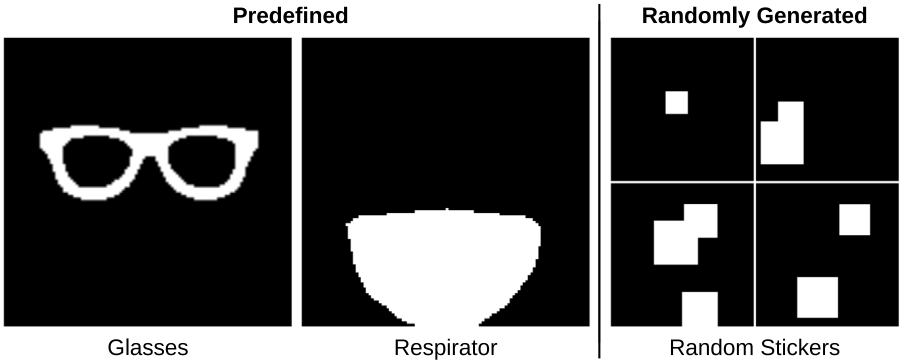


Fig. 6. Examples of the binary masks.

Table 4
The results of the impersonation attacks on different binary masks.

Target model	Method	Glasses			Respirator			Random Stickers		
		IR50	Effi-B0	Swin-T	IR50	Effi-B0	Swin-T	IR50	Effi-B0	Swin-T
IR50	PGDAP	62.20[a]	8.00	5.20	99.80[a]	27.80	31.60	**32.40**[a]	6.40	6.20
	TAP	17.40[a]	4.00	3.40	99.60[a]	25.20	35.40	16.20[a]	4.20	4.00
	EAP	**68.60**[a]	**17.60**	**5.60**	**100.00**[a]	**45.80**	**42.40**	30.00[a]	**10.80**	**6.40**
Effi-B0	PGDAP	5.80	**78.00**[a]	**4.80**	25.60	99.40[a]	24.80	4.00	**42.00**[a]	5.40
	TAP	1.20	18.80[a]	2.80	19.20	99.60[a]	28.20	1.60	19.20[a]	4.20
	EAP	**9.20**	75.40[a]	4.40	**40.60**	**100.00**[a]	**37.40**	**5.40**	37.40[a]	**6.20**
Swin-T	PGDAP	1.20	2.60	19.00[a]	22.60	23.00	95.00[a]	1.60	3.40	14.00[a]
	TAP	4.60	7.00	**83.20**[a]	27.20	26.20	**100.00**[a]	4.60	5.60	**45.20**[a]
	EAP	**18.00**	**23.60**	75.40[a]	**56.80**	**60.80**	**100.00**[a]	**10.20**	**15.80**	33.80[a]

[a] Indicates white-box attacks.

lower MCS for target-attack pairs than other services, suggesting that they has stronger defense mechanisms against adversarial patches. On the other hand, Huawei has relatively higher MCS for target-attack pairs than other services, indicating that it has a weaker discriminative ability between different identities.

Attacks on Different Binary Masks. Table 4 presents the results of impersonation attacks on various binary masks using the CelebA-HQ dataset. Fig. 6 provides visual examples of distinct binary masks. All other settings remain consistent with those configured for the attack on the backbone network.

Upon examination of Table 4, it is apparent that our approach outperforms alternative methods when applied to various masks. Nevertheless, it is important to note that, when compared to impersonation attacks using eye masks, our method does not achieve effective black-box impersonation on these masks. This observation underscores the critical role of meticulously designing a suitable binary mask to successfully execute an effective black-box impersonation attack.

Furthermore, Fig. 7 illustrates the visualization results of impersonation attacks using respirator binary masks. The evidence from this data indicates that the EAP attack method outperforms others in these specific impersonation attacks.

Table 5
Comparative experiment results on direct pasting of target facial patches, evaluated using mean ASR for backbones and heads, and MCS for Face++ and Tencent systems.

Method	Models		Systems	
	Backbones	Heads	Face++	Tencent
PASTE	39.24	46.15	66.21	30.37
EAP	**91.22**	**80.96**	70.25	34.58
EAP-M	–	–	**75.94**	**50.69**

Comparative Study with Pasting of Target Facial Patches. In Table 5, we detail a comparative study on the CelebA-HQ dataset, examining the impact of the “PASTE” method on target facial patches. The effectiveness of “PASTE” is compared with our advanced methods, EAP and EAP-M, the latter being an iteration based on a meta-ensemble attack strategy. The data highlights the superior performance of EAP and EAP-M over PASTE, particularly against commercial facial recognition systems like Face++ and Tencent, demonstrating the robustness of our methods.

Complementing this analysis, Fig. 8 provides a visual representation of the impersonation attacks executed using the PASTE method. It

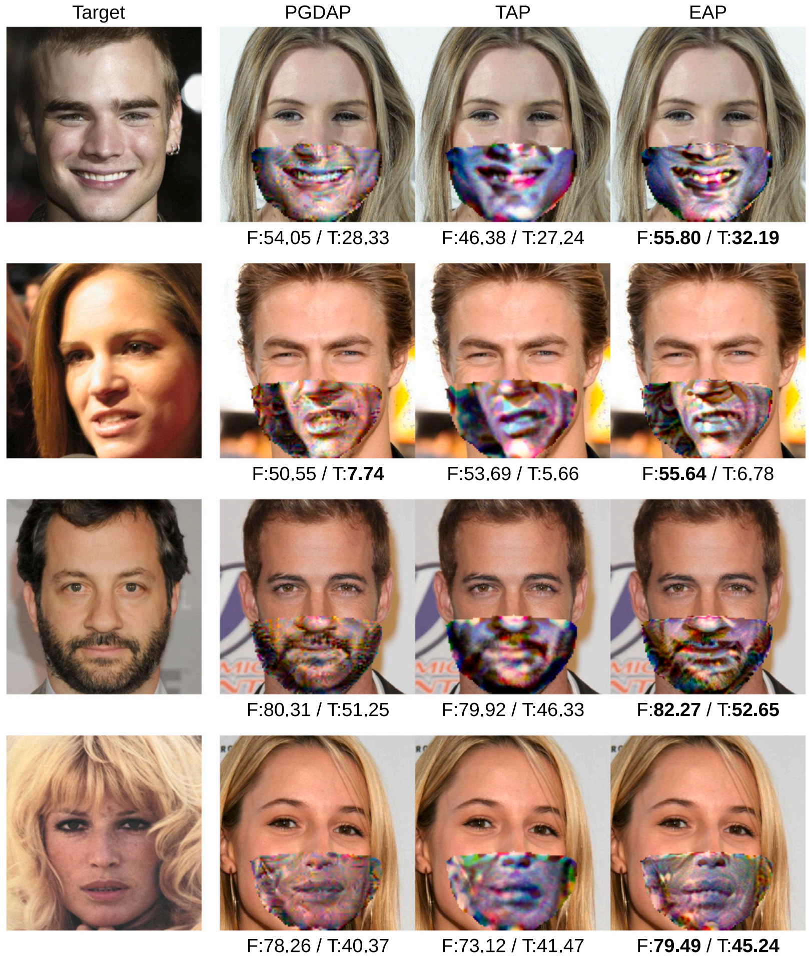


Fig. 7. Visualization results of impersonation attacks on respirator binary masks, with confidence scores displayed below each image. Notations (F: Face++, T: Tencent) indicate the source of each score. The target model in these attacks is IR50.

displays the confidence scores assigned by Face++ and Tencent for each impersonated image, offering a direct and quantitative insight into the method's ability to deceive these systems. The visual data reinforce the comparative findings, highlighting the relative effectiveness of the different methods in challenging the accuracy and reliability of contemporary facial recognition technologies.

Ablation Study. In Table 6, we report the results of an ablation study conducted to evaluate the effectiveness of EAP. The table compares different strategies used in this attack such as random similarity transformation (RST) strategy, image pyramid (IP) strategy and meta-ensemble attack (MEA) strategy. It can be seen that using all three strategies together yields better performance than when any one strategy is used alone. RST strategy improves the transferability of adversarial patches by increasing the diversity of input images. IP strategy further enhances the transferability by adapting to different scales and resolutions of input images. MEA strategy extracts more common gradient features from multiple face models and generates more robust adversarial patches. The combination of all three components achieves the best performance for black-box impersonation attacks on face recognition models and systems.

Sensitivity Analysis of Hyperparameters. In Fig. 9, we give the results of the sensitivity experiments for the hyperparameter β used to control the similarity transformation and the hyperparameter γ used to control the image pyramid levels.

In Fig. 9, we can see that an appropriate β value can improve the transferability of adversarial patches, while too large a value can lead to performance degradation. Furthermore, larger values of γ require more computational resources, making them less practical in real-world scenarios. Based on these findings, we recommend setting β between 0.1 and 0.2 and γ to 3 or 4 for optimal performance and efficiency.

Table 6
The results of the ablation study. "RST" means the random similarity transformation strategy, "IP" means the image pyramid strategy, and "MEA" means the meta-ensemble attack strategy. The results of the backbones and the heads use the mean ASR. The results of the systems use the mean MCS.

RST	IP	MEA	Backbones	Heads	Systems
✘	✘	✘	60.46	52.67	45.80
✔	✘	✘	87.85	76.56	54.45
✘	✔	✘	76.28	65.24	50.23
✔	✔	✘	**91.22**	**80.96**	**57.48**
✘	✘	✔	–	–	65.62
✔	✘	✔	–	–	66.63
✘	✔	✔	–	–	67.52
✔	✔	✔	–	–	**69.08**

4.3. *Physical-realizability experiments*

Simply succeeding in digital attacks does not necessarily mean they can be applied in the physical world. Physical attacks hold more practical value in real-world scenarios, compared to digital ones. To conduct experiments on physical attacks, the iPhone 11 Pro Max's 12MP front-facing camera was used for imaging, and we utilized a

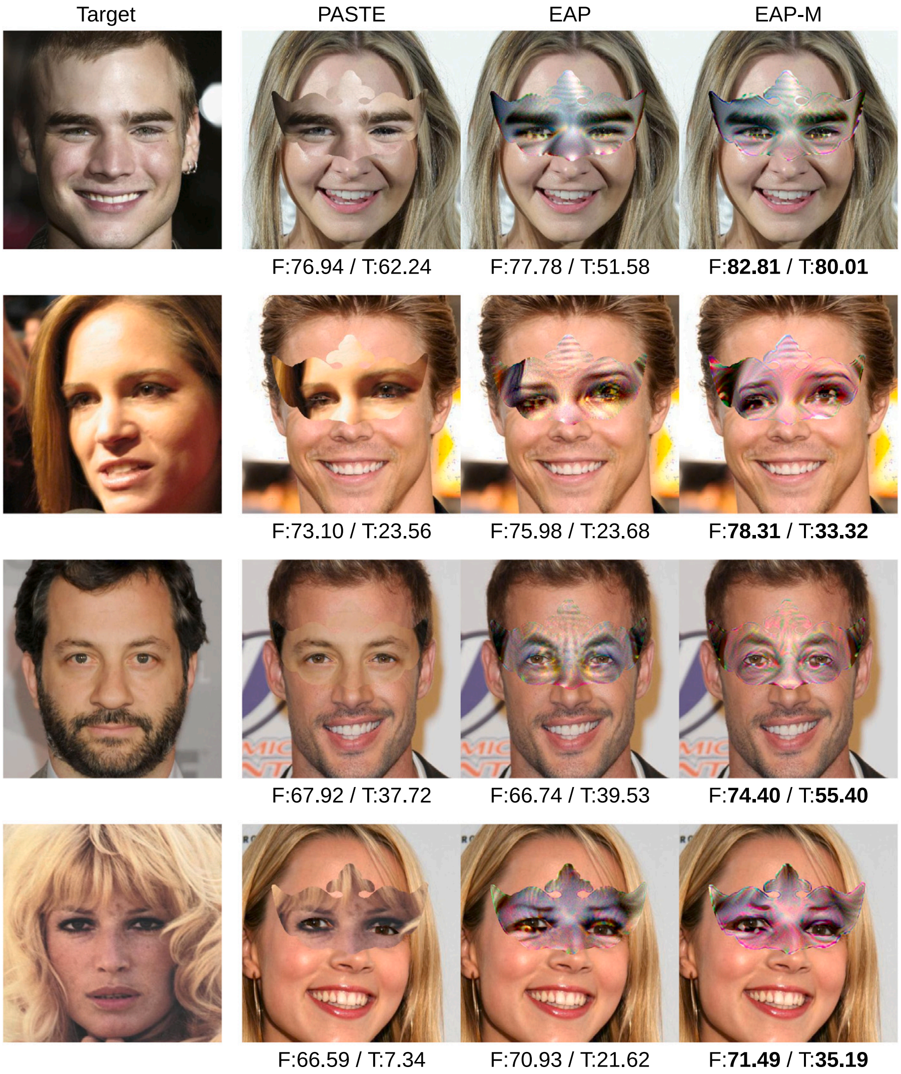


Fig. 8. Visualization results of impersonation attacks using PASTE, with confidence scores displayed below each image. Notations (F: Face++, T: Tencent) indicate the source of each score.

mobile and compact printer, the Canon SELPHY CP1300, to print adversarial patches. Fig. 10 shows the visualization results of our physical black-box impersonation attacks on five state-of-the-art commercial face recognition systems. The target and source remain the same as in Fig. 1. Compared to TAP, our EAP method achieves higher target-attack confidence scores on all five commercial systems and at all positions. At position ①, EAP-M effectively attacked all five commercial face recognition systems, achieving confidence scores exceeding 70%. Notably, even when the quality of the face was poor, EAP-M exhibited commendable attack performance at the long-range position ④. This shows that our method can achieve effective black-box impersonation attacks in both digital and physical environments.

5. Discussion

The development of EAP signifies an innovative step in adversarial attack methodologies for face recognition systems, particularly in physical environments. EAP's unique contribution is its practicality and effectiveness in real-world scenarios, a domain that has been less explored in the context of face recognition security.

Distinctive Approach: A key innovation of EAP is its ability to facilitate effective impersonation attacks using conveniently accessible mobile and compact printers. This practicality in implementation sets EAP apart from existing methods, as it significantly lowers the barrier to executing physical world attacks. By enabling the use of common printing devices, EAP demonstrates a novel approach to creating adversarial examples that are both accessible and effective in real-world scenarios.

Innovative Strategies: The introduction of random similarity transformation and image pyramid strategies is a novel contribution to the field. These strategies significantly improve the transferability and robustness of adversarial patches across different face recognition systems. Furthermore, the meta-ensemble attack strategy is a pioneering approach in utilizing common gradient features from multiple models, enhancing the success rate of impersonation attacks.

Practical Implications and Challenges: While EAP offers a new avenue for impersonation attacks, it also brings to light the practical challenges in designing and deploying adversarial examples in physical environments. The balance between effectiveness and non-detectability remains a critical area for further research.

Security and Ethical Considerations: The effectiveness of EAP raises significant concerns about the security of current face recognition systems and highlights the necessity for ongoing research into robust defense mechanisms. It also prompts a broader discussion on the ethical implications of adversarial research in the context of privacy and security.

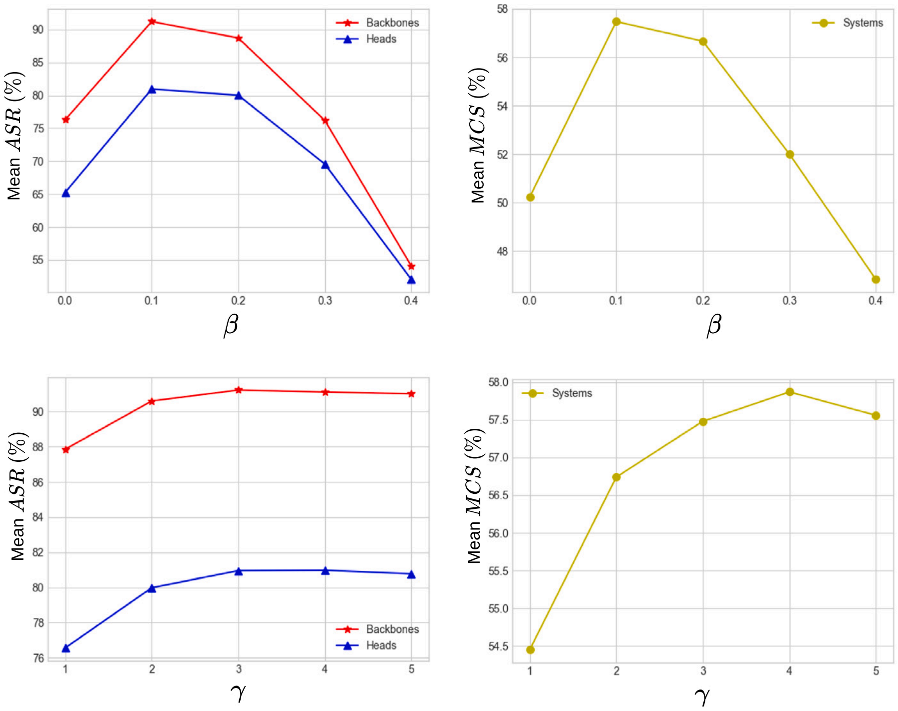


Fig. 9. The results of the β and γ sensitivity experiments.

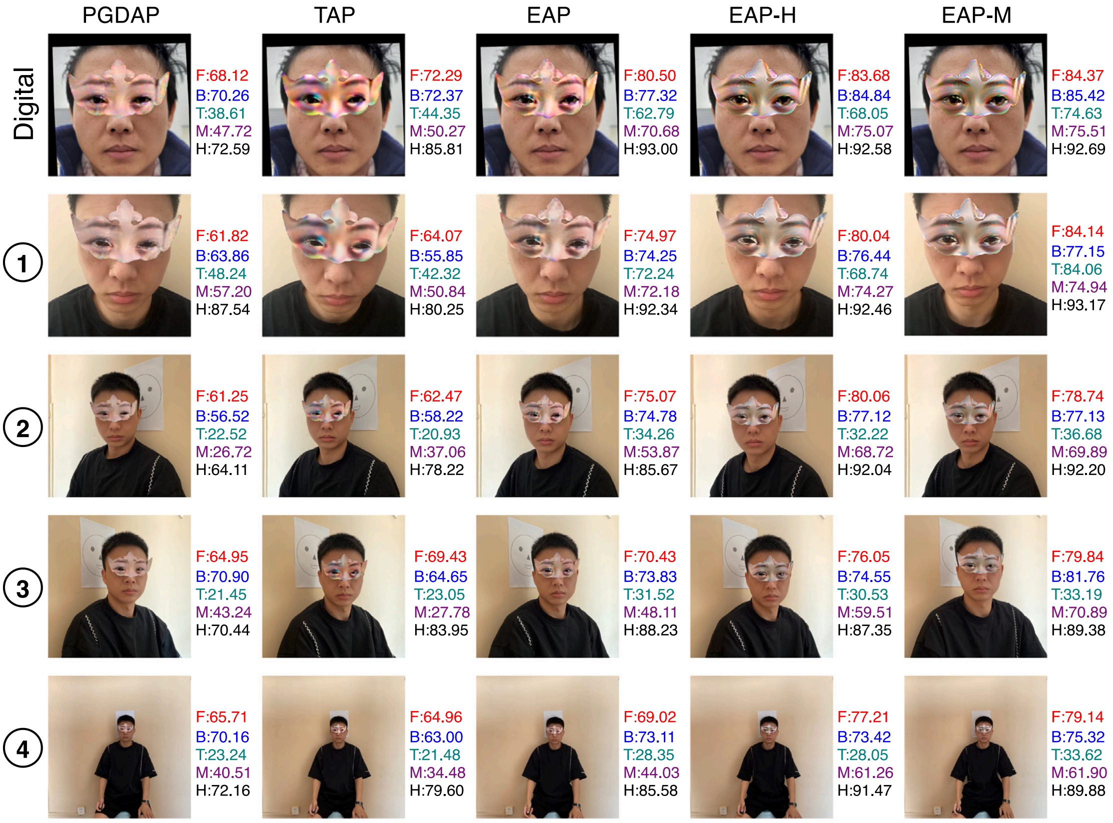

PGDAP
TAP
EAP
EAP-H
EAP-M
Digital
F:68.12 B:70.26 T:38.61 M:47.72 H:72.59
F:72.29 B:72.37 T:44.35 M:50.27 H:85.81
F:80.50 B:77.32 T:62.79 M:70.68 H:93.00
F:83.68 B:84.84 T:68.05 M:75.07 H:92.58
F:84.37 B:85.42 T:74.63 M:75.51 H:92.69
1
F:61.82 B:63.86 T:48.24 M:57.20 H:87.54
F:64.07 B:55.85 T:42.32 M:50.84 H:80.25
F:74.97 B:74.25 T:72.24 M:72.18 H:92.34
F:80.04 B:76.44 T:68.74 M:74.27 H:92.46
F:84.14 B:77.15 T:84.06 M:74.94 H:93.17
2
F:61.25 B:56.52 T:22.52 M:26.72 H:64.11
F:62.47 B:58.22 T:20.93 M:37.06 H:78.22
F:75.07 B:74.78 T:34.26 M:53.87 H:85.67
F:80.06 B:77.12 T:32.22 M:68.72 H:92.04
F:78.74 B:77.13 T:36.68 M:69.89 H:92.20
3
F:64.95 B:70.90 T:21.45 M:43.24 H:70.44
F:69.43 B:64.65 T:23.05 M:27.78 H:83.95
F:70.43 B:73.83 T:31.52 M:48.11 H:88.23
F:76.05 B:74.55 T:30.53 M:59.51 H:87.35
F:79.84 B:81.76 T:33.19 M:70.89 H:89.38
4
F:65.71 B:70.16 T:23.24 M:40.51 H:72.16
F:64.96 B:63.00 T:21.48 M:34.48 H:79.60
F:69.02 B:73.11 T:28.35 M:44.03 H:85.58
F:77.21 B:73.42 T:28.05 M:61.26 H:91.47
F:79.14 B:75.32 T:33.62 M:61.90 H:89.88

Fig. 10. The visualization results of physical black-box impersonation attacks on five commercial face recognition systems. The confidence scores are pasted to the right of each attack, (F:Face++, B:Baidu, T:Tencent, M:Microsoft, H:Huawei).

Future Research Directions: Future studies could focus on refining EAP's methodology to further minimize detectability while maintaining effectiveness. Exploring defense strategies against such adversarial attacks is also crucial for enhancing the security of face recognition systems.

EAP represents a step forward in understanding and demonstrating the vulnerabilities of face recognition systems, especially in physical settings, offering valuable insights for both offensive and defensive strategies in cybersecurity.

6. Conclusions

In conclusion, we have proposed an effective black-box impersonation adversarial patch attack method on face recognition, which we call EAP. Our approach includes a random similarity transformation strategy, an image pyramid strategy, and a meta-ensemble attack strategy to improve the transferability and generalization of adversarial patches. Our experiments demonstrate that EAP is a highly effective attack method against state-of-the-art face recognition models and commercial systems.

Importantly, our work highlights the vulnerability of current commercial face recognition systems, which can be easily compromised by attackers with access to real face images of target identities from social networks. These findings have significant implications for the security and privacy of individuals, and further research is needed to address these vulnerabilities and develop more robust defense mechanisms.

Overall, our proposed method has the potential to contribute to the development of more secure face recognition systems, as well as to enhance our understanding of the limitations and vulnerabilities of current models.

CRediT authorship contribution statement

Xiaoliang Liu: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Furao Shen:** Investigation, Project administration, Resources, Supervision. **Jian Zhao:** Methodology, Writing – review & editing. **Changhai Nie:** Resources, Supervision.

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.

Data availability

The datasets generated and analyzed during this study are available in the https://drive.google.com/drive/folders/1w7F2-PnXSXDJQti9x2Bsed5VDrAzmo3t?usp=drive_link.

Acknowledgments

This work was supported in part by the STI 2030-Major Projects of China under Grant 2021ZD0201300, and by the National Science Foundation of China under Grant 62276127.

References

[1] 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: ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 1528–1540.
[2] M. Sharif, S. Bhagavatula, L. Bauer, M.K. Reiter, A general framework for adversarial examples with objectives, ACM Trans. Priv. Secur. 22 (3) (2019) 1–30.
[3] D.-L. Nguyen, S.S. Arora, Y. Wu, H. Yang, Adversarial light projection attacks on face recognition systems: A feasibility study, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 814–815.
[4] S. Komkov, A. Petiushko, AdvHat: Real-world adversarial attack on ArcFace face ID system, in: International Conference on Pattern Recognition, ICPR, IEEE Computer Society, 2021, pp. 819–826.
[5] 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: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 7714–7722.
[6] Y. Zhong, W. Deng, Towards transferable adversarial attack against deep face recognition, IEEE Trans. Inf. Forensics Secur. 16 (2020) 1452–1466.
[7] L. Yang, Q. Song, Y. Wu, Attacks on state-of-the-art face recognition using attentional adversarial attack generative network, Multimedia Tools Appl. 80 (1) (2021) 855–875.
[8] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2014, arXiv preprint arXiv:1412.6572.
[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, ICLR, 2018.
[10] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: IEEE Symposium on Security and Privacy, SP, IEEE, 2017, pp. 39–57.
[11] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.-J. Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 15–26.
[12] A. Ilyas, L. Engstrom, A. Athalye, J. Lin, Black-box adversarial attacks with limited queries and information, in: Proceedings of International Conference on Machine Learning, ICML, PMLR, 2018, pp. 2137–2146.
[13] Y. Liu, X. Chen, C. Liu, D. Song, Delving into transferable adversarial examples and black-box attacks, 2016, arXiv preprint arXiv:1611.02770.
[14] Y. Dong, T. Pang, H. Su, J. Zhu, Evading defenses to transferable adversarial examples by translation-invariant attacks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 4312–4321.
[15] J. Kos, I. Fischer, D. Song, Adversarial examples for generative models, in: 2018 IEEE Security and Privacy Workshops, SPW, IEEE Computer Society, 2018, pp. 36–42.
[16] A. Zolfi, S. Avidan, Y. Elovici, A. Shabtai, Adversarial mask: Real-world universal adversarial attack on face recognition models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2022, pp. 304–320.
[17] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, C. Liu, Adv-makeup: A new imperceptible and transferable attack on face recognition, in: International Joint Conference on Artificial Intelligence, IJCAI, 2021.
[18] X. Wei, Y. Guo, J. Yu, Adversarial sticker: A stealthy attack method in the physical world, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[19] Z. Xiao, X. Gao, C. Fu, Y. Dong, W. Gao, X. Zhang, J. Zhou, 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.
[20] H. Jin, S. Liao, L. Shao, Pixel-in-pixel net: Towards efficient facial landmark detection in the wild, Int. J. Comput. Vis. (IJCV) (2021) http://dx.doi.org/10.1007/s11263-021-01521-4.
[21] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: IEEE Conference on Computer Vision and Pattern Recognition , CVPR, 2018, pp. 9185–9193.
[22] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, A.L. Yuille, Improving transferability of adversarial examples with input diversity, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 2730–2739.
[23] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.
[24] G. Seni, J.F. Elder, Ensemble methods in data mining: improving accuracy through combining predictions, Synth. Lect. Data Min. Knowl. Discov. 2 (1) (2010) 1–126.
[25] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of International Conference on Machine Learning, ICML, PMLR, 2017, pp. 1126–1135.
[26] Q. Sun, Y. Liu, T.-S. Chua, B. Schiele, Meta-transfer learning for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition , CVPR, 2019, pp. 403–412.
[27] M.A. Jamal, G.-J. Qi, Task agnostic meta-learning for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition , CVPR, 2019.
[28] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, ICLR, 2018.
[29] G.B. Huang, M. Mattar, T. Berg, E. 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.
[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[31] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[33] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[34] Y. Hu, X. Wu, R. He, Tf-nas: Rethinking three search freedoms of latency-constrained differentiable neural architecture search, in: European Conference on Computer Vision, Springer, 2020, pp. 123–139.

[35] D. Han, S. Yun, B. Heo, Y. Yoo, Rethinking channel dimensions for efficient model design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 732–741.

[36] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2736–2746.

[37] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: Making vgg-style convnets great again, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13733–13742.

[38] S. Chen, Y. Liu, X. Gao, Z. Han, MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices, in: Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings, vol. 10996, Springer, 2018, p. 428.

[39] X. Wu, R. He, Z. Sun, T. Tan, A light CNN for deep face representation with noisy labels, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2884–2896.

[40] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3349–3364.

[41] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: More features from cheap operations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1580–1589.

[42] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[43] X. Zhang, R. Zhao, Y. Qiao, X. Wang, H. Li, Adacos: Adaptively scaling cosine logits for effectively learning deep face representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10823–10832.

[44] H. Liu, X. Zhu, Z. Lei, S.Z. Li, Adaptiveface: Adaptive margin and sampling for face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11947–11956.

[45] F. Wang, J. Cheng, W. Liu, H. Liu, Additive margin softmax for face verification, IEEE Signal Process. Lett. 25 (7) (2018).

[46] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 4690–4699.

[47] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified perspective of pair similarity optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6398–6407.

[48] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, F. Huang, Curricularface: adaptive curriculum learning loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5901–5910.

[49] Q. Meng, S. Zhao, Z. Huang, F. Zhou, Magface: A universal representation for face recognition and quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14225–14234.

[50] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, T. Mei, Mis-classified vector guided softmax loss for face recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07, 2020, pp. 12241–12248.

[51] D. Zeng, H. Shi, H. Du, J. Wang, Z. Lei, T. Mei, NPCFace: Negative-positive collaborative training for large-scale face recognition, 2020, arXiv preprint arXiv:2007.10172.

[52] Face++, 2022. https://www.faceplusplus.com. (Last Accessed on 11 February 2023).

[53] Baidu, 2023. https://ai.baidu.com/tech/face. (Last Accessed on 11 February 2023).

[54] Tencent, 2023. https://cloud.tencent.com/product/facerecognition. (Last Accessed on 11 February 2023).

[55] Microsoft, 2023. https://azure.microsoft.com/en-us/services/cognitive-services/face. (Last Accessed on 11 February 2023).

[56] Huawei, 2023. https://www.huaweicloud.com/product/face.html. (Last Accessed on 11 February 2023).

[57] J. Wang, Y. Liu, Y. Hu, H. Shi, T. Mei, Facex-zoo: A pytorch toolbox for face recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3779–3782.

[58] T.B. Brown, D. Mané, A. Roy, M. Abadi, J. Gilmer, Adversarial patch, 2017, arXiv preprint arXiv:1712.09665.

[59] D. Deb, J. Zhang, A.K. Jain, Advfaces: Adversarial face synthesis, in: IEEE International Joint Conference on Biometrics, IJCB, 2020.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Neural Inf. Process. Syst. (NeurIPS) 32 (2019).

Xiaoliang Liu obtained his B.S. degree in communication engineering from Fujian University of Technology, Fujian, China. He is currently pursuing a Ph.D. in the Department of Computer Science and Technology at Nanjing University, Jiangsu, China. His research primarily focuses on adversarial machine learning, artificial neural networks, and computer vision.

Furao Shen (Member, IEEE) completed his B.Sc. and M.Sc. degrees in mathematics at Nanjing University, Nanjing, China, in 1995 and 1998, respectively. He earned his Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2006. Currently, he holds the position of Full Professor of the School of Artificial Intelligence at Nanjing University. His research interests revolve around neural computing and robotic intelligence.

Jian Zhao (Senior Member, IEEE) earned his B.S. degree from Nanjing University, Nanjing, China, his M.Sc. degree from the Hamburg University of Technology, Hamburg, Germany, and his Dr. Sc. degree in electrical engineering from the Swiss Federal Institute of Technology (ETH) Zurich, Switzerland. From 2010 to 2015, he served as a Research Scientist at the Institute for Infocomm Research, A*STAR, Singapore. Currently, he is an Associate Professor at the School of Electronic Science and Engineering, Nanjing University. His research interests encompass deep neural networks, mathematical optimization, and wireless communication networks. Dr. Zhao's achievements include the Dengfeng Scholars Program of Nanjing University in 2015, IEEE Globecom 2008 Best Paper Award, and the 2009 Chinese Government Award for Outstanding Self-Financed Students Abroad.

Changhai Nie holds the position of a professor in software engineering at the Department of Computer Science and Technology at Nanjing University, where he is affiliated with the National Key Laboratory for Novel Software Technology. His primary research interests lie in the fields of software testing and search-based software engineering, with a particular focus on combinatorial testing, search-based software testing, and methods for comparing and combining software testing techniques, among others.