Got My “Invisibility” Patch: Towards Physical Evasion Attacks on Black-Box Face Detection Systems

Duohe Ma, *Member, IEEE*, Junye Jiang, Xiaoyan Sun, *Member, IEEE*, Zhimin Tang, Zhenchao Zhang, Kai Chen, *Member, IEEE*, and Jun Dai, *Member, IEEE*

***Abstract*—Modern face detection (FD) systems have demonstrated remarkable performance in identifying human faces, primarily via Deep Neural Networks (DNNs). However, these DNN-driven models exhibit inherent susceptibility to adversarial attacks, posing significant risks for intentional face obfuscation from detectors. Such obfuscation can serve both malicious purposes (e.g., evading surveillance systems) and benign objectives (e.g., protecting personal privacy). Previous studies have developed techniques to compromise the effectiveness of various FD models, yet these adversarial attacks are largely confined to the digital domain—e.g., by applying adversarial perturbations to digital input images—or demand prior knowledge of the target FD systems. In this paper, we introduce a novel framework for evading black-box face detection (FD) systems in real-world scenarios. The proposed method relies on the *Expectation over Attention* (EoA) algorithm, which generates the *Public Attention Heat Map* (PAHM) by fusing attention mechanisms across an ensemble of publicly available FD models. Our evaluation results demonstrate that EoA outperforms state-of-the-art (SOTA) methods in white-box settings and demonstrates strong cross-model transferability in black-box scenarios, effectively evading FD systems across smartphones, laptops, and surveillance cameras.**

***Index Terms*—Face detection, adversarial attack, black-box.**

I. Introduction

FACE detection (FD) is a fundamental computer vision technique that identifies the presence of human faces in images or videos. It plays a critical role in various applications such as face alignment [1], [2], [3], face recognition [4], [5], [6], facial expression analysis [7], and face tracking [8]. Recent advances in deep learning, particularly in Deep Neural Networks (DNNs) [9], have significantly enhanced the accuracy and efficiency of FD. Hence, this technique has been widely adopted in many scenarios of our daily life, such as video surveillance [10], entry access management [11], unmanned store [12] and payment authorization [13]. As an illustrative example, the global response to the COVID-19 pandemic has led to the deployment of face recognition systems at transportation hubs to help register visitors and monitor body temperatures. The seamless operation of these functions relies heavily on the proper functioning of the FD module.

While DNNs have significantly enhanced the performance of FD systems, they are well known to be vulnerable to adversarial attacks [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. Adversaries can introduce subtle alterations to input data to deliberately mislead DNN models into producing incorrect outputs. Previous research has studied adversarial attacks across various domains, such as computer vision [25], natural language processing [26] and reinforcement learning [27]. FD adversarial attack is an evasion technique that manipulates input data (e.g., images or videos) by injecting imperceptible perturbations to induce failures in FD systems, such as missing genuine faces (false negatives) or misclassifying non-facial regions as faces (false positives). Researchers have developed a range of attacks designed to allow adversaries to conceal their faces from detectors by strategically applying carefully-crafted patches [28], [29], [30], [31], [32]. This category of attacks known as FD evasion attacks poses inherent security risks in critical applications such as video surveillance and criminal investigation. On the other hand, it’s important to note that these attacks also have the potential to serve as a means of safeguarding human privacy by thwarting unauthorized or malicious attempts to collect individuals’ facial information.

However, the existing FD evasion attacks have limitations in the following aspects. First, some attacks were implemented only within the digital realm [28], [30] by manipulating the input images directly. These attacks assume that the adversary has access to input face images or is capable of executing unrestricted pixel-level alterations to these images. However, applying these attack techniques directly to real-world FD systems is ineffectual: 1) in most cases the input images are not readily accessible to attackers for alteration; 2) impacting images at pixel-level through physical channels is challenging due to the unpredictable physical conditions [33] or the inherent flaws present in devices (e.g. printers and cameras) when manipulating the

Received 30 October 2023; revised 6 June 2025; accepted 7 June 2025. Date of publication 23 June 2025; date of current version 4 November 2025. The work of Duohe Ma, Junye Jiang, Zhimin Tang, Zhenchao Zhang, and Kai Chen was supported by the National Natural Science Foundation of China under Grant 62472418.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the IRB-equivalent review committee at the Institute of Information Engineering, Chinese Academy of Sciences.

Duohe Ma, Junye Jiang, Zhimin Tang, Zhenchao Zhang, and Kai Chen are with the Institute of Information Engineering of Chinese Academy of Sciences, Beijing 100085, China (e-mail: maduohe@iie.ac.cn; jiangjunye@iie.ac.cn; tangzhimin@iie.ac.cn; zhangzhenchao@iie.ac.cn; chenkai7274@iie.ac.cn).

Xiaoyan Sun and Jun Dai are with the Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609 USA (Corresponding author: Jun Dai) (e-mail: xsun7@wpi.edu; jdai@wpi.edu).

Digital Object Identifier 10.1109/TDSC.2025.3582290

1545-5971 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

pixels [34]. Second, some attacks require what is often referred to as "white-box" access to the target system. In simpler terms, it assumes the adversary knows the target models (algorithms, network architecture, parameters) in advance, based on which they can craft the patches [29]. Nevertheless, this requirement frequently proves impractical, since companies specializing in face detection are typically reluctant to disclose their algorithmic structure and parameter details. This reluctance primarily arises from concerns related to intellectual property rights and security considerations. It's important to note that in reality, most face detectors operate as "black-box" systems, meaning that the internal functional modules (such as pre-processing, DNN, and post-processing) are not disclosed to the public. The absence of comprehensive knowledge about the target system can significantly diminish the feasibility of such attacks.

Due to above limitations, very few solutions can effectively and efficiently conduct evasion attacks to the "black-box" FD systems in the physical world [29], [31], [35], [36]. To bridge the gap, our paper proposes a novel solution to evade face detection in black-box FD applications. Our approach can tackle two challenges: 1) the input face images may not be accessible to attackers and thus cannot be manipulated; and 2) specific knowledge (e.g. models, algorithms) about the target FD system may not be available.

To address the first challenge, we introduce an approach that employs physical-world adversarial patches applied to facial regions, thereby eliminating the need for direct manipulation of input images in the digital domain. However, creating such adversarial patches usually needs prior knowledge of the specific FD systems. To further tackle this second challenge, we use an algorithm called *Expectation over Attention* (EoA), for the attack. Specifically, we devise a method to generate adversarial perturbations based on the attention *heat maps* of the FD model, rather than its gradients. Using an attention-based loss function, it shifts the FD detection model's focus away from the face, causing it to lose focus and enabling face detection evasion. To ensure our attack works with no specific knowledge about the target system, we propose to leverage an ensemble of publicly available FD models. From this ensemble, we generate the *Public Attention Heat Map* (PAHM), a term we crafted to provide a comprehensive representation of where general FD models focus their attentions. Adversarial patches generated from PAHM exhibit feasibility and generality in attacking a variety of FD systems powered by different algorithms and models. Our proposed adversarial patches exhibit cross-model effectiveness across multiple FD systems (see Section IV-D). This breakthrough eliminates the constraints of previous evasion attacks which were confined to a specific FD system, and removes the attack requirement of knowing the intricate functional modules within target FD systems.

To comprehensively assess the effectiveness, generality, and robustness of our attack, we conducted experiments towards FD systems across a range of devices. These devices included four different smartphones (iPhone, Samsung, Xiaomi, and VIVO), one laptop (Mechrevo X3-S PC), and a surveillance camera equipped with an infrared temperature measurement module and a face detection module. Our experimental procedures adhered to ethical guidelines for human subject research (equivalent to the Institutional Review Board (IRB) approval system in the United States).[1] The devices in our experiments run different black-box FD applications, such as built-in camera apps, BeautyCam, Alipay, among others, and we had no access to specific details about these FD apps. Our physical experiments yielded compelling results, demonstrating that our adversarial patches could effectively and successfully disable the face detection functions in these systems. The diverse range of face appearances, represented by our carefully selected 17 participants, highlights the robustness of our algorithm. Moreover, through contrast experiments, we found that the evasion effect remained consistent even when participants wore hats or glasses. We also validated our approach by comparing it with other attack methods [29], [31] described in the existing literature, and our results showed a significant improvement in the attack success rate. Beyond evaluating attack performance on commercial black-box models, we conducted rigorous cross-model validation to assess adversarial transferability. Confidence scores from multiple unseen recent FD models were measured under clean versus adversarial conditions. The observed confidence degradation empirically confirms that EoA-generated patches effectively transfer across diverse FD models.

To the best of our knowledge, our approach is the first to effectively evade black-box FD systems in the physical domain, across multiple models and without requirement for specific knowledge of the target models. The major contributions of this paper are as follows: 1) We presented a new approach to conduct evasion attacks on black-box FD systems in the physical realm, without relying on the specific knowledge of the target systems. 2) We introduced the Public Attention *Heat Map* to evade a variety of FD systems with different detection models and algorithms. 3) We conducted experiments on different black-box FD systems across a range of devices and validated the effectiveness of our approach.

The findings and outcomes presented in our paper shed important insights to the field of face detection (FD) research and development: 1) our method serves as a means to assess the security and robustness of FD models deployed in benign real-world applications. When our generated adversarial patches successfully breach a target FD system in use, it indicates the need to enhance its security measures; 2) in an era where many businesses collect users' facial information without their explicit consent, our method assumes a pivotal role in protecting human privacy by preventing unauthorized or malicious collection of individuals' facial information.

The rest of this paper is structured as follows. In Section II, we give the background of face detection, state the problem with threat model, and summarize the related works with limitations. In Section III, we provide the details of our attack methodology. In Section IV, we present our evaluation results. We discuss the limitations and future works in Section V, and conclude in Section VI.

The source code and video demonstration are publicly shared via GitHub to support community research[2].

[1] We obtained explicit permission and consent from seventeen participants before proceeding with experiments. All face images used in this paper were not only authorized but also blurred for anonymization.

[2] https://github.com/JJuny123/EOA/tree/main

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

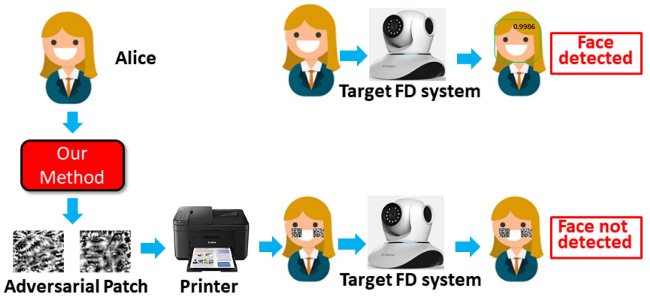


Fig. 1. When Alice wears our adversarial patches (printed by a laser printer) on her cheeks, she can escape the face detection system in the real world.

II. Background and Preliminaries

A. Face Detection (FD)

FD is a object detection technique specifically emphasizing the recognition of human faces. Traditional approaches leverage hand-craft features to identify faces [37], [38], [39], [40], including local features [41], boosting algorithms [42], [43], cascade structures [2], [44], [45], [46] and deformable part models [47].

Benefiting from the development of deep learning, Convolutional Neural Networks (CNNs) have been utilized to achieve more effective face detection. These approaches are mainly built upon three types of network architectures: (1) Some two-shot detectors [48], [49], [50], [51], [52] are based on RCNN (Region-based CNN) [53], in which they first identify the region proposals from the image, and then classify whether each region is a face or not; (2) Some other two-shot methods [54], [55], [56], [57] are based on Faster-RCNN [58], in which they adopt the Region Proposal Networks (RPN) to generate the initial region proposals by a set of anchor boxes and predict the confidence score and bounding box offset for each anchor box, and then refine the proposals to obtain the final output; (3) Some single-shot approaches [59], [60], [61], [62], [63], [64] are based on SSD (a Single-Shot Detector) [65], in which they locate a set of anchor boxes, and predict the confidence score and bounding box offset.

As is commonly known by the community, deep learning models are vulnerable to adversarial attacks [15]. By introducing subtle perturbations to the input samples, the model can be misled to wrong decisions. Instead of directly changing the digital input images, this paper aims to develop adversarial attacks against DNN-based face detectors by creating physical face patches. Our attack does not rely on the knowledge of target FD systems, and is featured by its general effectiveness in evading a variety of detection models, regardless of their network architectures or algorithms.

B. Problem Statement and Attack Requirements

The primary objective of this paper is to create an effective attack capable of evading a variety of black-box FD systems. We consider vision-based FD systems that are designed to recognize human faces. Our approach is to generate adversarial patches and put them on a person's face, so that the target FD system can no longer detect that face (Fig. 1). It's important to note that an adversarial patch's dimensions must be significantly smaller than those of human faces so that wearing it can fool human observers. Such FD evasion attacks have been investigated in previous works [28], [29], [30], [31], [32], enabling adversaries to potentially circumvent surveillance or criminal investigations. Going beyond prior works, we consider the following practical requirements for our attack.

1) Physical Attack: Our objective is to develop and deploy adversarial attacks in the physical domain for FD systems. Unlike digital attacks that directly manipulate input pixel values, adversaries in this scenario must design printable adversarial patterns, physically fabricate them as wearable accessories, and strategically position them on facial regions. When captured by surveillance cameras, these patterns can disrupt the facial detection process during imaging. The multi-stage conversion process, from digital design to physical materialization (printing) and back to digital representation (camera capture), introduces substantial non-linear distortions in pixel space. This physical-to-digital transformation pipeline fundamentally alters attack constraints: The inevitable information loss during materialization (printer resolution limitations) and re-digitization (camera sensor noise, lighting variations, perspective distortions) eliminates the feasibility of end-to-end gradient propagation. Consequently, traditional iterative optimization methods for attack refinement become prohibitively inefficient compared to purely digital attack scenarios, necessitating new approaches for robust physical attack generation.

2) Attacking Black-Box Systems: We target widely used FD systems in the real world, to demonstrate our crafted attack and its severity and practicality. These systems typically exhibit greater complexity comprising various modules, including pre-processing, deep learning models, and post-processing stages. The models employed in these systems tend to be resilient against adversarial attacks or noise. More importantly, these real-world systems are typically "black-box" in nature, meaning they remain partially or completely opaque to potential adversaries. Consequently, adversaries do not have any insights into the inner workings of the detection mechanisms, including crucial details such as model algorithms, structures, data processing methods, and training datasets. This lack of information presents significant challenges when it comes to devising effective attack strategies.

3) General and Robust: Our framework prioritizes two key objectives: target-agnostic generality and environmental robustness. It achieves cross-model evasion using a universal adversarial perturbation ensemble, eliminating the need for model-specific patch generation. Additionally, it maintains effective detection evasion across diverse real-world conditions, including varying illumination, poses, occlusions (e.g., hats), and distances.

C. Existing Attacks and Their Limitations

Efforts have been made in the research community to study methods for compromising FD. However, these approaches are limited by certain constraints.

Initially, attacks were proposed to undermine DNN-based facial detectors within digital environments [28], [30].

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

These tactics leveraged traditional gradient-based [14] or optimization-based approaches [25] to generate adversarial perturbations, either in white-box or black-box settings. However, transitioning these solutions to physical world scenarios poses challenges. First, the unavailability of design rationales and implementation details in black-box FD system makes it impractical to acquire precise or approximate gradients for generating perturbations. Second, the effectiveness of attacks in the physical realm can significantly drop due to real-world distortions caused by various environmental factors such as distances and angles, lighting conditions, printing quality and camera resolutions [66].

In a subsequent effort, evading FD system in the physical dimension was explored in [29], which targeted the Multi-Task Cascaded Convolutional Neural Networks (MTCNN) model [67]. In this approach, the adversary computes gradients and generates perturbations against the white-box MTCNN model, using the Expectation over Transformation (EoT) [68] technique. These perturbations can then be printed as patches and affixed to a person's face, enabling them to evade face detection. The effectiveness of this method in extending the attack from digital to physical is largely contingent on the adversary's white-box access to the target model and the comprehensive understanding towards the model. However, when the adversary lacks knowledge of the victim system, as is the case in this study, the attack will fail. Additionally, it's worth noting that the adversarial patch generated in this context [29] is not general and hence inapplicable to other systems. Zhou et al. [31] designed a more generalized method to break the black-box FD systems in the physical realm. However, their success rate remained low due to the limited transferability between the publicly available model and the target system.

In addition to evading face detection, a variety of studies have been conducted for devising adversarial attacks against general object detection models, particularly in real-world scenarios. For instance, Chen et al. [69] targeted the Faster-RNN model with the goal to fool it into mis-detecting stop traffic signs based on the EoT technique. Eykholt et al. [34] accomplished similar attacks against the YOLOv2 model by incorporating patterns into stop signs. Thys et al. [70] considered intra-class variation to generate patches capable of concealing individuals from the YOLOv2 model. Zhao et al. [71] introduced a set of innovative techniques to enhance the robustness of physical attacks. These attacks are also confined ones that work only for a specific target system, and lacks generality and robustness for working towards varied target systems in the real world.

III. Methodology

A. Overview

As analyzed above, most existing attacks on facial detection operate in the digital domain, which usually achieve evasion by adjusting input images at the pixel level. However, applying these attack techniques directly to real-world FD systems is impractical, as adversaries usually lack access to the digital images captured by such systems, and impacting images at pixel-level through physical channels is also challenging due to the unpredictable physical conditions or the inherent flaws present in devices (e.g. printers and cameras).

Few studies have explored physical-domain attacks in face detection evasion, with [29] and [31] being among the few that fall into this category. These works study adversarial attacks relying on gradient signals to generate perturbations.

Although gradients indicate which direction reduces the loss the most, they have two key issues: first, they optimize for local decision boundaries of specific models, resulting in poor transferability to black box systems with architecture or training differences. Second, they lack semantic interpretability and often disrupt non-salient regions that humans consider irrelevant. Consequently, these attacks suffer from low success rates or limited transferability when applied to unknown black-box models.

To overcome these limitations, we strategically choose to generate adversarial perturbations from the attention *heat map* rather than gradients, as it provides a more comprehensive representation of where the FD model focuses its attention. We introduce a novel methodology, **Expectation over Attention (EoA)**, to attack black-box FD systems in the physical domain. This solution is founded on two key concepts.

First, inspired by [72], we generate adversarial perturbations from the attention *heat map* instead of the gradients. Gradients indicate which direction loss decreases the most, while the attention *heat map* (`AHM`) provides a much more comprehensive representation of where a deep learning model concentrates its attention when making decisions [73]. `AHM` also contains rich information exploitable by the adversary to compromise the input and fool the model. Based on this understanding, we propose a loss function predicated on the attention *heat map*, which can be leveraged to make the target model lose its focus on crucial facial regions within the input image.

Second, we propose to create an ensemble of open-source FD models and compute attention *heat maps* for each of them. An average Public Attention *Heat Map* (`PAHM`) can be subsequently derived from these individual *heat maps*, which collectively encapsulates the characteristics of various FD models. It is on the basis of `PAHM` that we then generate adversarial patches.

By targeting the attention distribution instead of the original gradient, our method systematically disrupts the salient regions essential for face alignment across different models. This shift from local gradient optimization to global attention distortion significantly enhances transferability.

Another limitation of existing face detection evasion attacks is their reliance on white-box assumptions. In contrast, our method demonstrates strong cross-architecture transferability in decision-based black-box settings. Through benchmark evaluations on six face detection models excluded from the EoA training pipeline, we empirically validate that EoA achieves a high success rate against black-box models. Please refer to Section IV-D for a demonstration of our model's robustness and transferability, where we analyze changes in face detection confidence before and after the attack on six latest face detection models: *RetinaFace* [74], *CenterFace* [75], *YOLO8Face*, *SCRFD* [76], *DBFace*, and *UltraFace*.

Fig. 2 illustrates the workflow of our devised attack, which can be divided into preparation and an iterative process that consists

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

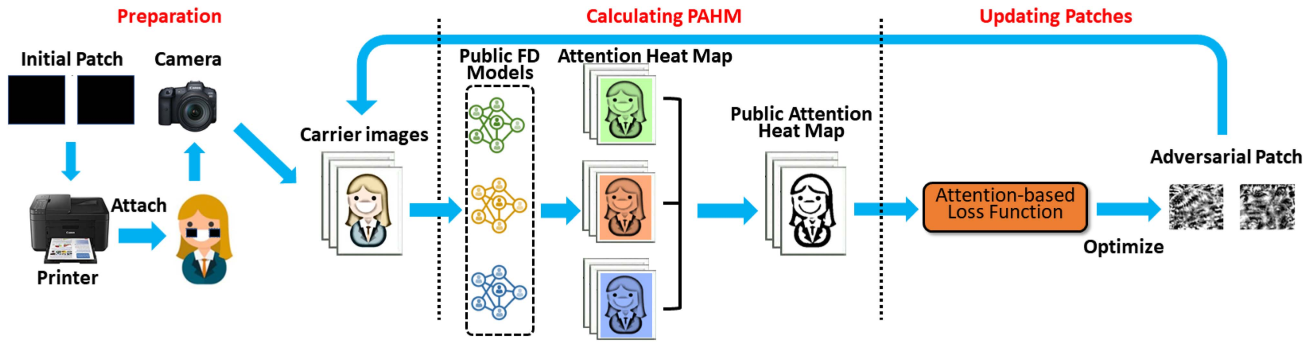
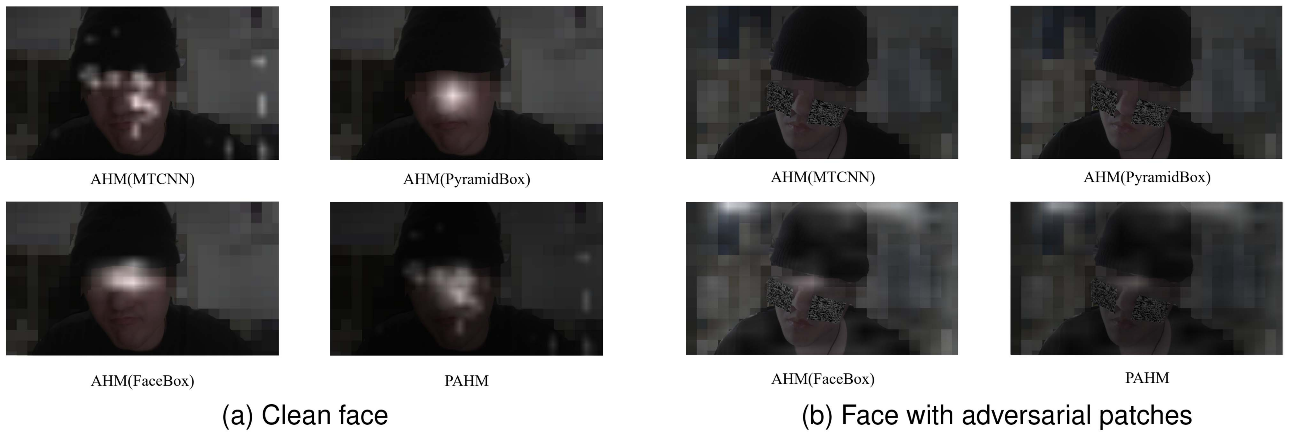


Fig. 2. The workflow of generating adversarial patches against varied FD systems.



(a) Clean face

(b) Face with adversarial patches

Fig. 3. Attention *heat maps* for three public FD models and the averaged public attention *heat map* (PAHM).

of two stages. *Preparation* takes squares of a pre-determined size as initial patches, *attaches* printed patches onto the cheeks of the person to evade FD, and takes photos, generating a number of carrier images (explained in Section III-B). The following Stage I, named "*Calculating PAHM*" constructs the corresponding PAHM. To achieve this, we affix adversarial patches from the previous iteration onto the person's cheeks, capture photos, input these images into an ensemble of publicly available FD models, and subsequently calculate the PAHM. The subsequent Stage II, named "*Updating Patches*" revolves around fine-tuning adversarial patches based on PAHM, with the objective to pinpoint patches that can divert the FD system's attention away from facial areas of the person. This objective is probed through optimizing an attention loss function. Stage I and Stage II are iterated for a certain number of times until satisfactory patches are obtained.

Below we describe the details of each step, and Algorithm 1 depicts the entire workflow.

Algorithm 1: Our EoA Attack Against FD Systems.

```
Input: $a$: accumulator; $K$: number of optimization iterations; $N$:
    number of photos; $\mu$: attention rate; $\epsilon$: learning rate
Output: $\delta x$: adversarial patch content
/* Preparation */
Print out two patch squares and attach them to the tester's cheek
Take $N$ photos of the tester's face with different conditions
Mark the patch edges in the photos to obtain $\{x_i | i \in N\}$
$\delta x$ = black pixels
$a = 0$
for $k \in \{1, ..., K\}$ do
    /* Stage I: Calculating PAHM */
    for $i \in \{1, ..., N\}$ do
        Project $\delta x$ to each $x_i$ to obtain $\{\hat{x_i} = x_i + \delta x | i \in N\}$
        Feed $\hat{x_i}$ to an ensemble of $M$ public FD models
        Compute $\texttt{PAHM}_i$ from Equations 1 and 2
    end
    /* Stage II: Updating Patches */
    Calculate $\mathcal{L} = L_{\texttt{PAHM}} + \gamma L_{TV}$ from Equations 3 – 5
    $a = \mu \times a + \nabla_{\delta x}\mathcal{L} / \|\nabla_{\delta x}\mathcal{L}\|$
    $\delta x = \texttt{Clip}(\delta x - \epsilon \times \texttt{sign}(a))$
end
```

B. Preparation

In this paper we craft our adversarial patches as two squares, each measuring 5 cm × 6 cm, attached to both sides of the individual's cheek. The size and placement of the patches adhere to the same parameters as outlined in the previous research [29]. Notably, a trade-off exists between the size of the patch and the effectiveness of evading detection. Adversaries may seek to reduce the dimensions of printed adversarial patches to improve their concealment from human observation. However, the reduction in patch size inevitably comes at the cost of a lower success rate in evading detection. Currently the primary focus of the research community is to deceive DNN models with the escape success rate taking precedence. The patch size reported in this

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

paper represents the optimal compromise weighing both facets of the trade-off. To make adversarial patches imperceptible to human observation, an alternative solution is to camouflage them within facial or head coverings. In Section IV-H, we demonstrate a successful strategy to generate and camouflage adversarial patches as part of a medical mask, reaching its imperceptibility while preserving effectiveness in evading FD.

Fig. 3 and many other figures throughout this paper illustrate an individual (namely tester) wearing the patches. Prior to initiating the attack procedure, we print two squares with a laser printer and affix them to the tester's cheek. It's important to note that, at this stage, we are solely concerned with capturing the patch regions and do not yet consider the patch content, which will be addressed in subsequent steps. We follow to use a camera to take photographs of the tester wearing the two patch squares. Various environmental or technical factors can affect the effectiveness of these patches, including lighting conditions, patch rotation, patch size, as well as distortion noise or blurriness introduced by the camera. Based on this understanding, we capture photos at different distances (30 cm, 60 cm, and 90 cm), angles ($-30°$, $0°$, and $30°$), and lighting conditions (bright light, dim light), to make our adversarial patches robust. Hence, we take a set of photographs of the tester, and use N to denote the quantity of the photographs captured. The value of N is a factor contributing to the attack success rate, as well as the complexity of the patch generation. The selection of N will be discussed in Section IV.

For each of the N photographs, we then delineate sides of the two patch squares through manual marking. Each square can be identified by nine distinct points. This gives us a set of N *carrier images* denoted as $\{x_i | i \in N\}$. These carrier images will be used throughout the attack workflow, as the patch content undergoes iterative updates.

C. Stage I: Calculating PAHM

This stage is to calculate the PAHM for the N carrier images of a given person (i.e., the tester in our illustration). We perform the following steps to compute PAHM at each iteration.

1) Adding Patch Content to Carrier Images: The first step is to introduce digital patch content, denoted as δx, into the set of N carrier images. More precisely, during the first iteration, all pixels of δx are initialized as black. In each subsequent iteration, we utilize δx obtained from the previous iteration. We then proceed to map δx onto the marked patch regions within each of the carrier images, yielding a set of N *perturbed images*, denoted as $\{\hat{x_i} = x_i + \delta x | i \in N\}$.

2) Computing Attention Heat maps and PAHM: We then compute the PAHM of the set of N perturbed images, a task that requires computation of attention *heat maps* corresponding to individual FD models in the ensemble, as explained in Section III-A. The ensemble is the enabler of the generality (and applicability to black-box targets) of our approach, for which we choose to include M widely-used FD models in the ensemble, all of which are either publicly accessible or open-source. These models vary in terms of their structures and algorithms, and they may differ from the target FD system.

We feed all the perturbed images into each of the M publicly available FD models. For each image $\hat{x_i}$ and model m_j, we denote $y_{i,j}$ as the highest probability that indicates the presence of a face in that image. Then the corresponding attention *heat map*, denoted as $\texttt{AHM}_{i,j}$, can be computed as follows using the Grad-CAM method [77]:

$$\alpha^r_{i,j} = \frac{1}{Z}\sum_p\sum_q \frac{\partial y_{i,j}}{\partial A_i^{p,q,r}}$$

$$\texttt{AHM}_{i,j} = \texttt{ReLU}\left(\sum_r \alpha^r_{i,j} A^r_i\right) \tag{1}$$

where $A_i^{p,q,r}$ is the pixel value at position p, q of the r-th channel's feature map, Z is the number of pixels for each channel, and $\alpha^r_{i,j}$ is the weight of the r-th channel's feature map. A^r_i is the r-th channel's feature map, and the attention *heat map* $\texttt{AHM}_{i,j}$ is computed as the weighted sum of the feature maps from all the channels with the Rectified Linear Unit (ReLU) function. This step is essential as negative values in the *heat map* lack meaningful interpretations.

For each perturbed image $\hat{x_i}$, we can now compute the public attention *heat map* $\texttt{PAHM}_i$ by averaging the attention *heat maps* of all the selected models:

$$\texttt{PAHM}_i = \sum_{j=1}^{M} \texttt{AHM}_{i,j} \tag{2}$$

D. Step II: Updating Adversarial Patches

Based on the PAHM obtained in Stage I, we proceed to craft adversarial patches. The PAHM essentially quantifies the significance of different regions within an image concerning the face detection task. The activation area in the *heat map* is expected to closely align with the actual facial features, enabling the detector to recognize the existence and location of the face [72]. Hence, to evade FD, our strategy is to reduce the concentration of the model on the facial area, causing the detector to fail in face recognition. Fig. 3 illustrates the attention *heat maps* from three publicly available FD models (MTCNN [67][3], PyramidBox [60][4] and Facebox [78][5]) and the averaged PAHM, respectively for a clean (i.e., unmodified) face (Fig. 3(a)) and for an adversarial-patched face (Fig. 3(b)). Our observation reveals that FD models primarily concentrate their attention on the person's face, while our introduced patches effectively divert this attention away.

We then continue to compute the variance (the average of the squared differences from the mean) of all the pixels within the PAHM and limit or clip this value to the range of $(0, 1)$. Our objective is to minimize this variance to ensure that the attention of the face detector is evenly distributed across the entire image, preventing it from concentrating on and detecting the facial region. This problem can be formulated as the following loss function, where $L_{\texttt{PAHM}}$ denotes the thermal loss function, N denotes the number of input images, Clip is the Python function to limit all pixel values in PAHM to the range (0, 1), $\|\cdot\|$ is the norm calculation function, $Avg(\cdot)$ is the mean calculation

[3] https://github.com/edosedgar/mtcnnattack/tree/master/mtcnn
[4] https://github.com/EricZgw/PyramidBox
[5] https://github.com/610265158/faceboxes-tensorflow/tree/tf1

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

function, and $\sum$ is the summation function.

$$L_{\texttt{PAHM}} = \frac{1}{N}\sum_{i=1}^{N} \|\texttt{Clip}(\texttt{PAHM}_i - \texttt{Avg}(\texttt{PAHM}_i))\| \quad (3)$$

Furthermore, to avoid abrupt color transitions (for visual coherence) on the generated patches, we incorporate the total variation loss [79] to regulate the pixel values within the patches. This constraint aids in creating patches that appear more natural and visually smooth, as depicted below, where δ represents the digital patch content, and $p_{i,j}$ denotes the pixel value at position (i, j).

$$L_{TV} = \sum_{p_{i,j}\in\delta x} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2} \quad (4)$$

Therefore, our goal is to find the δx that can minimize the following loss function, where γ is a hyper-parameter to balance the thermal loss and the total variation loss.

$$\mathcal{L} = L_{\texttt{PAHM}} + \gamma L_{TV} \quad (5)$$

This can be effectively addressed by using widely-used optimization techniques. As a representative example, we opt to employ the MI-FGSM algorithm [80] to iteratively generate the optimal δx. Specifically, we update the patches based on the gradient sign and then feed them to Stages I and II for the recomputation of PAHM and $\mathcal{L}$ until we get the ultimate set of patches. The final patches can then be printed and employed to launch an attack against the target FD system.

E. Implementation

The system design has been implemented as a prototype named *EOA*, consisting of approximately *2000* lines of Python code. The key custom classes in the codebase include: the *TrainMask* class, which is used to implement the training algorithm; the *LossManager* class, which is used to manage loss functions; the *PatchPartTF* class is used to control adversarial patch blocks; the *ImageTF* class, which is used to acquire calibration data from the input face image for adversarial sample generation; the *PatchTF* class, which is used for patch initialization and other operations within the patch areas; and the *PatchManager* class, which is used to save input face images and adversarial sample patches.

The processor used for training is an Intel(R) Xeon(R) Platinum 8369HC CPU. The TensorFlow version employed is 1.10.0. The training time is approximately *15* minutes, based on training parameters specified in the next section.

IV. Evaluation

To substantiate the efficacy and feasibility of our proposed attack method, we conducted a series of experiments and ablation studies, aiming to provide a comprehensive assessment.

A. Experimental Configurations

Preparing the carrier images: As pointed out in Section III-B, to make our adversarial patches robust, we need to take environmental factors into consideration. And thus we capture photos under various conditions, including three different distances (30 cm, 60 cm, and 90 cm), three angles (−30°, 0°, and 30°), and two lighting scenarios (bright light, dim light). Specifically, for each tester, we acquire a set of $N = 8$ photographic samples featuring black-and-white checkerboard-patterned patches affixed to cheeks as carrier images. Parameter N was empirically determined through systematic experimentation, which showed that exceeding this threshold ($N > 8$) results in diminishing returns in attack success rate while significantly increasing both training temporal and computational costs.

TABLE I
Attack Success Rate of Our Attack and Existing Works

Target FD model	Method	Distance			
		30cm	60cm	90cm	Average
Light-DSFD	[29]	4.5%	15.5%	1.5%	7.17%
	[31]	71.5%	76.0%	69.5%	72.33%
	Ours	**78.0%**	**97.5%**	**73.5%**	**83.00%**
Yoloface	[29]	1.5%	20.5%	19.5%	14.0%
	[31]	80.5%	78.5%	66.0%	75.00%
	Ours	**82.5%**	**83.5%**	**69.5%**	**78.50%**

Bold values show the best results in the given experimental scenario.

TABLE II
Attack Success Rate of Our Attack With Different Numbers of Public Models

Target FD model	M	Distance			
		30cm	60cm	90cm	Average
Light-DSFD	1	34.5%	45.5%	36.5%	38.83%
	2	62.5%	85.5%	69.0%	72.33%
	3	**78.0%**	**97.5%**	**73.5%**	**83.00%**
Yoloface	1	35.5%	39.5%	32.5%	35.83%
	2	74.5%	78.0%	63.5%	72.00%
	3	**82.5%**	**83.5%**	**69.5%**	**78.50%**

Bold values show the best results in the given experimental scenario.

Generating adversarial patches: We need to select M publicly accessible FD models to generate the PAHM. It is worth noting that incorporating more public models can enhance attack transferability but also entail additional computational costs. Our empirical evaluation shows that $M = 3$ (i.e., using just 3 of these models) already yields satisfactory results. This choice is grounded on empirical assessment, which confirms that $M = 3$ yields better results than $M = 1$ or 2. Notably, as M increases, the efficacy improvement decreases. Therefore, setting $M = 3$ (i.e., using three FD models) strikes a balance, providing satisfactory performance and efficacy while avoiding additional computational overhead. Further experiments involving random model selection affirm that once the value of M is established, varying combinations of FD models does not substantially affect performance (their outcomes consistently align). Details about model selection and ablation analysis can be found in Section IV-E and Table II.

The three open-sourced FD models are: MTCNN [67], PyramidBox [60] and Facebox [78]. The MTCNN model has been widely embraced within the research community due to its smaller size and faster training speed. Our decision to incorporate it aligns with established practices in the literature. In contrast, Facebox, despite also being compact and efficient, offers a distinctive perspective. Additionally, we include PyramidBox (a larger model that imposes a slightly higher training burden), which emerged as the 2018 winner of the WIDER FACE Competition. Such choices of these models are driven by a balance between potential computation overhead and the diversity they offer among FD models. It’s important to note that our approach is adaptable to include other FD models as alternatives. However, for the purposes of this paper, we focus on these three commonly used FD models as a representative combination to illustrate the attack.

To generate the adversarial patches using Algorithm 1, we configure the number of optimization iterations as $K \geq 2000$. This specific value entails an exponential increase in computational resources, and the choice of 2000 has been determined through empirical experimentation to reach a balance between the desired efficacy and the associated computational cost. In alignment with experimental settings in related work, we initialize the learning rate ϵ at 60/255 and allow it to gradually decay to 1/255. The attention rate μ transitions from 0.9 to 0.99 during the optimization process.

Baselines: As introduced in Section II-C, a number of previous studies [28], [29], [30], [31] have developed attacks to evade FD models.

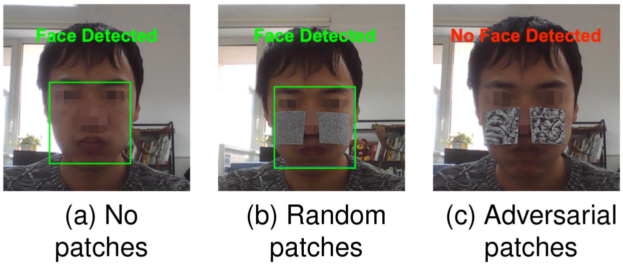
However, [28], [30] primarily focus on launching attacks in the digital realm, and hence their methods are not applicable to our specific attack context in the physical domain. Therefore, we select the approaches presented in [29], [31] as our benchmark references for comparisons, given that they are more closely aligned with our research objectives.

Target FD Systems: To comprehensively assess the generality and robustness of our attack method, we conduct tests across a diverse set of facial detection (FD) models and commercial systems. Specifically, our testbed devices include four smartphones (Samsung S10 5 G, iPhone XR, Xiaomi Redmi K20 Pro, Vivo X21), a Mechrevo X3-S laptop, and a surveillance camera with the infrared temperature sensors. Our evaluation encompasses the following FD models and software: 1) face detection with the functionality embedded into the camera applications of the smartphones and laptops; 2) commercial applications such as Beauty Camera B612 and Mobile Face Payment; 3) prominent FD models within the research community, including Light-DSFD and Yoloface. We use the HP LaserJet MFP M227 FDN printer to print the generated adversarial patches.

B. Attacking Black-Box FD Systems

Our experiments provide compelling evidence of the effectiveness of our proposed method in evading black-box FD systems. We emphasize that the black-box nature of the target systems is ensured through our strict adherence to the principle and practice of refraining from accessing any camera parameters, including model and algorithm specifics, of the target device to be attacked. This is particularly pertinent when dealing with commercial equipment, where such specifics are intentionally withheld from us. We also make sure the laptop camera utilized for data acquisition (i.e., photo capture) is distinct from any target device’s camera.

For each experiment, we implement and compare three cases: 1) the tester presents a clear face; 2) the tester wears patches randomly-generated based on various distributions; 3) the tester wears our adversarial patches. It’s worth noting that we have generated ten different types of random patches, each containing a classical random texture. The experimental results consistently show that random patches do not exhibit a natural evasion capability. These ten random patch variations and their ineffectiveness in FD evasion have been illustrated in Fig. 12. Green boxes indicate that faces were detected even with the random patches. Please note that we blurred the faces appeared in this paper to protect the experiment participants’ privacy. The original figures used in the experiments for FD were not blurred.



(a) No patches
(b) Random patches
(c) Adversarial patches

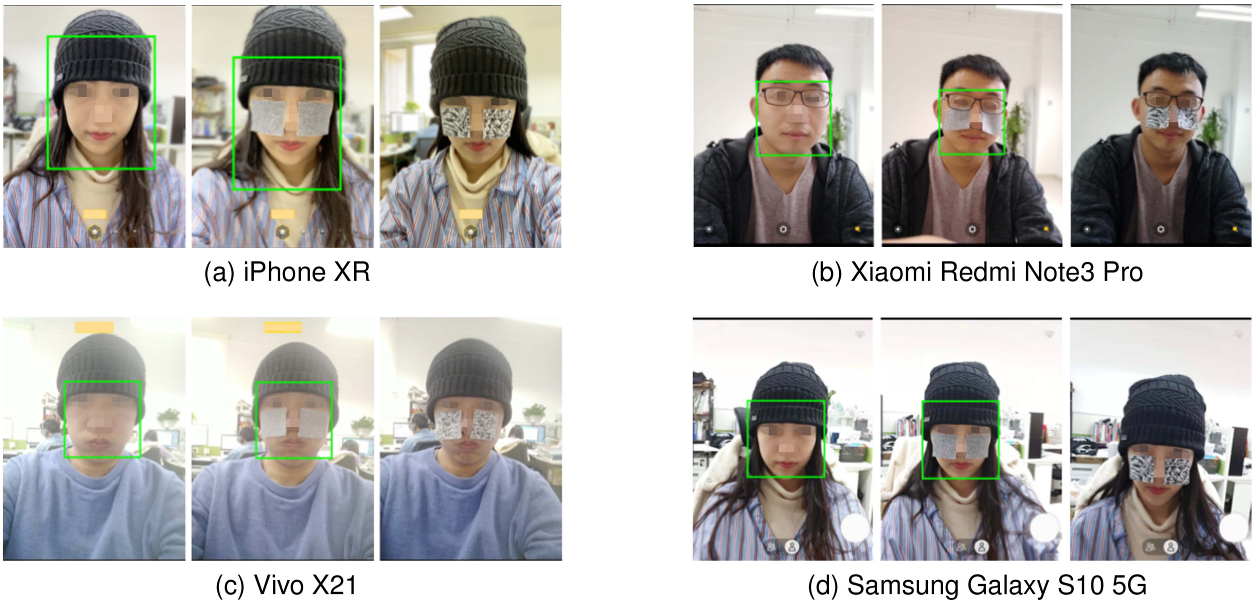
Fig. 4. Attacking the camera app of Windows 10. Faces were blurred to protect experiment participants’ privacy. Original figures used for FD experiments were not blurred.

Attacking the camera app of Windows 10: The Windows 10 operating system (OS) has a built-in camera application designed to capture images of users when they are positioned in front of the computer. It is equipped with an FD algorithm to detect human faces and correspondingly enhance picture quality, e.g., adjusting the brightness levels. An example of the application’s functionality is illustrated in Fig. 4(a), where the FD algorithm successfully detects the user’s face and highlights it with a green bounding box.

To launch attacks we ask the testers to wear patches and take photos using the laptop’s built-in camera app. The results are presented in Fig. 4(b) and (c), corresponding to scenarios involving random and adversarial patches, respectively. It is noticed that the random patches do not prevent the FD from detecting faces, as they fail to conceal critical facial features such as the eyes, nose, and mouth. In contrast, our intentionally crafted adversarial patches demonstrate a distinct capability. These patches enable the testers to evade detection successfully, as they effectively divert the model’s attention away from the facial region. These findings hold true across various testers and diverse conditions, yielding consistent results.

Attacking the camera apps of smartphones: Similarly, we test the efficacy of our attack against the built-in camera apps in four distinct smartphones. For each smartphone, we select the “portrait mode” in the camera app settings. The app will run an FD algorithm to identify potential facial features and mark faces (if found) with bounding boxes. For each smartphone, our attack is evaluated across the same three scenarios: when the tester wears no patches, when the tester wears random patches, and when the tester wears adversarial patches. Fig. 5 shows the representative detection results, in which each sub-figure corresponds to a particular smartphone and each contains comparative results under the three cases.

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

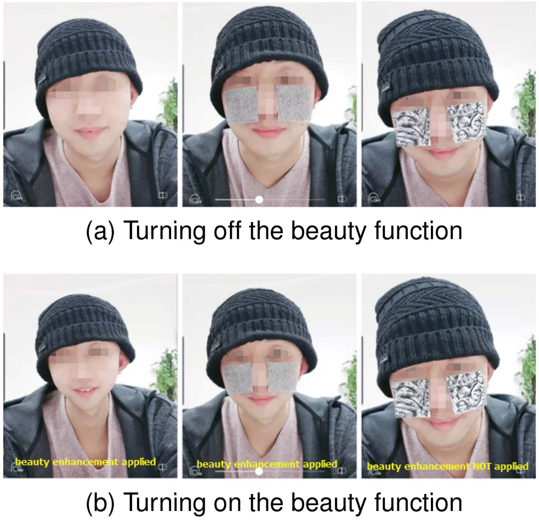

(a) iPhone XR

(b) Xiaomi Redmi Note3 Pro

(c) Vivo X21

(d) Samsung Galaxy S10 5G

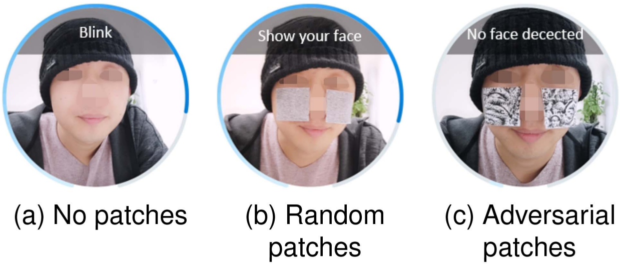
Fig. 5. Attacking the camera apps of smartphones. Green boxes indicate faces detected.


beauty enhancement applied
beauty enhancement applied
beauty enhancement NOT applied

(a) Turning off the beauty function

(b) Turning on the beauty function

Fig. 6. Attacking the B612 app (a photo/video editor).

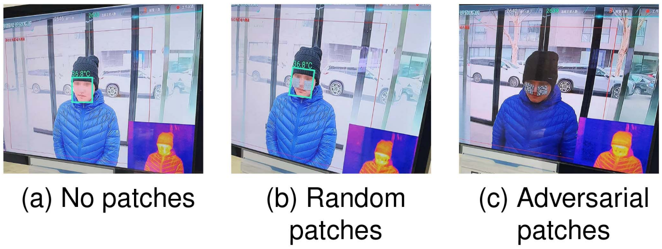


(a) No patches

(b) Random patches

(c) Adversarial patches

Fig. 7. Attacking Alipay app: (a) Tester asked to blink, (b) tester asked to tune the face angle, and (c) no face detected.


(a) No patches

(b) Random patches

(c) Adversarial patches

Fig. 8. Attacking the commercial surveillance camera.

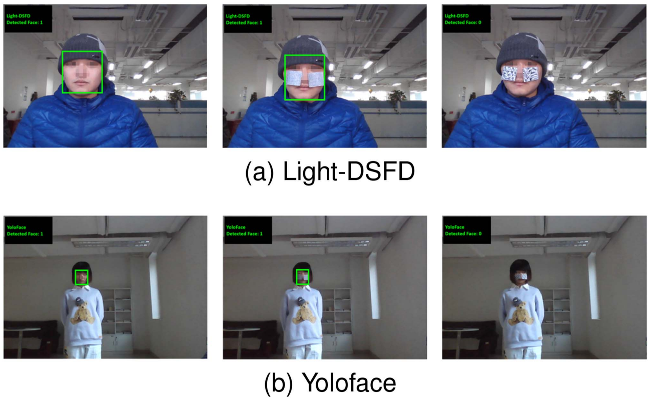
Our findings align well with the observations made in the laptop case: the FD algorithms exhibit the capability to detect faces, both in the absence of patches and when random patches are worn by the testers. However, our adversarial patches consistently prove effective in concealing the testers' faces from the smartphone cameras.

It's also noteworthy that our evaluation factors in various environmental variables, including lighting conditions, viewing angles, distance between the camera and the subject, as well as the presence of hats or glasses. The successful attacks using our approach, as shown in many results throughout this paper such as Figs. 8 and 9, span a wide spectrum of lighting conditions, viewing angles, and distances. Fig. 13 further shows that our attack is consistently effective in evading FD regardless of the size, shape, or color of the hats worn by different attackers.

Attacking commercial apps: We further consider two commercially available smartphone applications equipped with FD capabilities: 1) B612, which is a versatile photo and video editor, offering various beauty enhancements and filtering options; 2) Alipay Face Payment, which ranks among the most widely used payment apps globally. We evaluate our attack against the FD algorithms in these applications employing the same settings as the above experiments.

Fig. 6 presents the results obtained from attacking B612. The way that B612's beauty function works is that it will firstly perform the face detection, and then apply the beauty enhancements once faces are detected. The beauty function can

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.



(a) Light-DSFD

(b) Yoloface

Fig. 9. Attacking the research-based FD models.

be enabled or disabled in the application. In this experiment we explored two distinct scenarios: one with the beauty function disabled (Fig. 6(a)) and the other with the beauty function enabled (Fig. 6(b)). In both scenarios, when the tester did not wear any patch or wore random patches, the FD module in this application successfully identified the face. In this case, since beauty function was enabled in Fig. 6(b), the application further applied the beauty enhancements to the detected face by adjusting features like skin complexion and facial contour, as observed in the first two pictures in Fig. 6(b). However, in the last picture of 6(b), when the tester wore the adversarial patches, B612 could no longer detect faces and thus did not apply the beauty enhancements, even with the beauty function enabled. That's why the actual facial features in the last picture of Fig. 6(b) remained the same as the faces in Fig. 6(a) where beauty function was disabled.

Fig. 7 shows the results obtained when attacking the Alipay app. In Fig. 7(a), the tester does not wear any patch, and the app can recognize it is a face and prompts the message "please blink" to verify the presence of a live person. In Fig. 7(b), the tester wears random patches on his cheek, and the app can still detect the face and respond with the message "please tune the face angle", as a portion of the face is blocked by the patches. In Fig. 7(c), the tester wears the adversarial patches, and the app says "no face detected", indicating the FD failure. This confirms the effectiveness of our attack approach.

Attacking a commercial surveillance camera: In response to the global COVID-19 pandemic, many public spaces and buildings have deployed surveillance systems featuring infrared temperature sensors and FD cameras. These systems are intended to measure the face and body temperature of all visitors, however privacy intrusions have been reported in various locations around the world. In light of these concerns, we further evaluate our attack against the FD module in such a surveillance system using the same experimental settings as previously executed.

Fig. 8 shows the detection results in the context of surveillance cameras. Our observations in this scenario align with those outlined earlier: the FD algorithm effectively detects faces, both when the tester wears no patches and wears random patches. In contrast, our adversarial patches continue to demonstrate consistent effectiveness in concealing the tester's face from the surveillance camera's view. In the system that we test the infrared temperature sensor becomes non-operational when FD fails to identify the face.

We acknowledge that surveillance scenarios, such as border inspection, could benefit from human observation in conjunction with FD detection. The adversarial patches that we craft may be noticeable to human observers. Reducing the patch's size and camouflaging it (e.g., as a QR code that may represent a link to merchandise websites, which is a common marketing practice today) within facial or head coverings can provide a solution to further evade human detection. In Section IV-H, we demonstrate a successful strategy to generate and camouflage adversarial patches as part of a medical mask, to make the attack more imperceptible while preserving effectiveness in evading FD models. In this paper, our primary emphasis is on unmanned scenarios, with a central focus on studying methods to physically deceive various DNN models, giving priority to the escape success rate.

Attacking the research-based FD models: We also employ our adversarial patches to target two well-known models widely utilized in FD research: Light-DSFD [81] and Yoloface [82]. We avoid accessing any of the design or implementation specifics of the two models (i.e., treating them as black-box systems) and conduct our tests using the Mechrevo X3-S laptop equipped with a built-in camera as our testing platform. In Fig. 9, you can find examples of the detection results from attacking these two models. When testers wear our crafted patches, they are able to effectively bypass the face detectors, whereas wearing no patches or random patches does not yield the evasion effect.

C. Comparisons With Baselines

As explained in Section IV-A, two previously proposed FD evasion attacks [29], [31] can be used as baseline to compare with our methods. We select the Light-DSFD [81] and Yoloface [82], as representative target FD models to attack, and employ our generated patches to assess the effectiveness of these attacks.

To provide greater detail, our comparative study is conducted at three distinct distances (30 cm, 60 cm, and 90 cm) between the tester and the camera. We choose these specific distances to investigate and analyze how each distance impacts the success rate of attacks. We have also integrated the evaluation of other environmental factors, such as lighting conditions and viewing angles, along with the distance consideration. For each distance, we capture a total of 200 patched face images at various times of the day and different viewpoints (as detailed in Section IV-A). The attack success rate is calculated as the percentage of patched images that successfully evade detection by the target model.

Table I presents our EoA attack results in contrast to the two selected baseline methods [29], [31], across various target models and distances. It is evident that our EoA attack consistently outperforms the baseline methods. Notably, at the distance of 60 cm, our attack stands out with a significantly higher and more impressive attack success rate (i.e., 97.5%) compared to the baseline methods. This effectiveness is remarkable, especially considering that our approach can simultaneously subvert

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

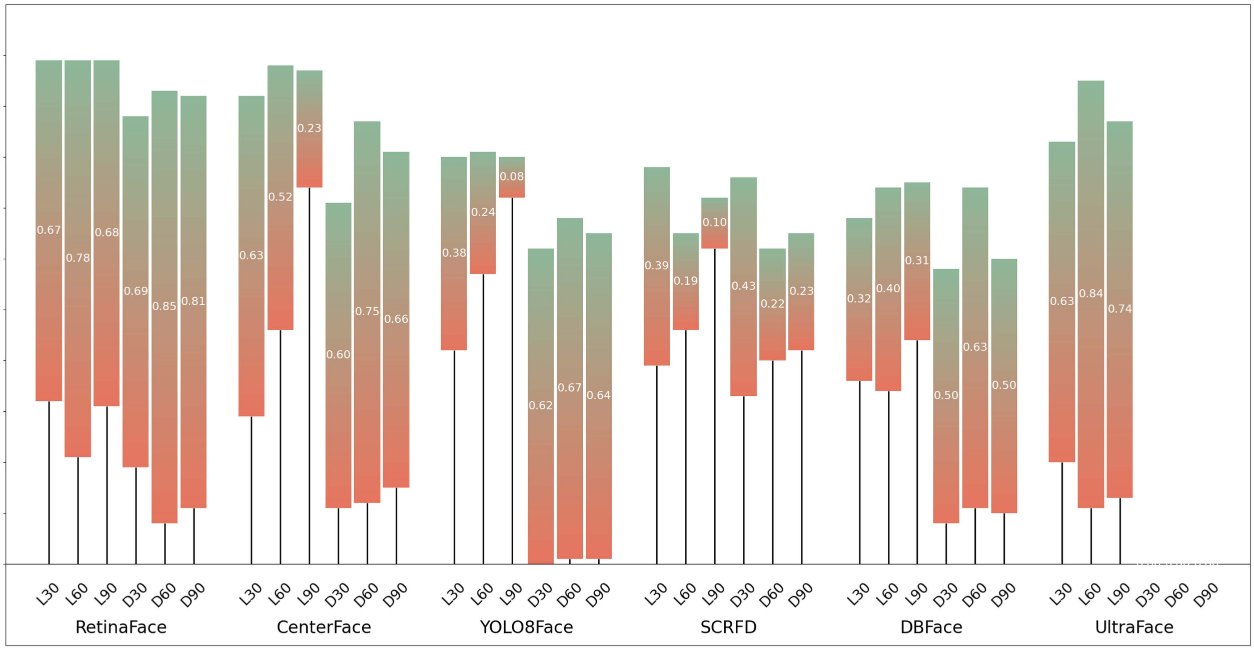


Fig. 10. Confidence reduction across different FD models due to EoA adversarial attack.

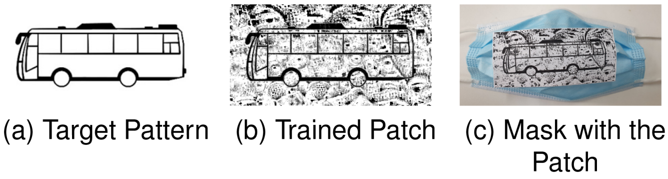

(a) Target Pattern (b) Trained Patch (c) Mask with the Patch

Fig. 11. Generating adversarial patches on medical masks.

multiple black-box FD models, a capability unmatched by any other existing methods.

D. Transferability and Robustness

The face detection (FD) models share a common attention mechanism, with its core logic relying on the feature set of the facial region. Our approach, EOA, generates misleading texture or edge information through adversarial perturbations, triggering shared underlying feature extraction biases across different models and causing a cross-model attention shift. Due to the nature of EOA, even when trained on earlier open-source FD models, it remains effective in evading relatively newer FD systems. Furthermore, since the perturbation design of EOA does not rely on specific model gradients but instead on the interpretability of attention distribution, it demonstrates strong robustness in real-world scenarios.

To demonstrate the robustness and transferability of the EOA algorithm, we selected an additional set of six newer FD models, *RetinaFace*[6], *CenterFace*[7], *YOLO8Face*[8], *SCRFD*[9], *DBFace*[10] *and UltraFace*[11], for testing. To clearly illustrate the evasion effect of EOA, we compared the confidence levels of these FD models on faces without perturbation versus those with adversarial interference. The change in confidence intuitively shows that the EOA algorithm effectively evades even the latest FD models, despite being trained on earlier open-source FD models. It is important to note that this study specifically evaluates the confidence calibration of the model's native outputs. In commercially deployed black-box FD systems, additional facial quality assessment modules are often used to filter captured images, reducing false positives and negatives in the FD pipeline. In such settings, the effectiveness of EOA is further enhanced.

We captured testers' photos at distances of 30 cm, 60 cm, and 90 cm from the camera under both dark and light conditions, with 50 photos per group. During each session, the tester adopted multiple postures, including turning, tilting and looking straight, ensuring their eyes were always visible. For consistency, postures were standardized across all groups.

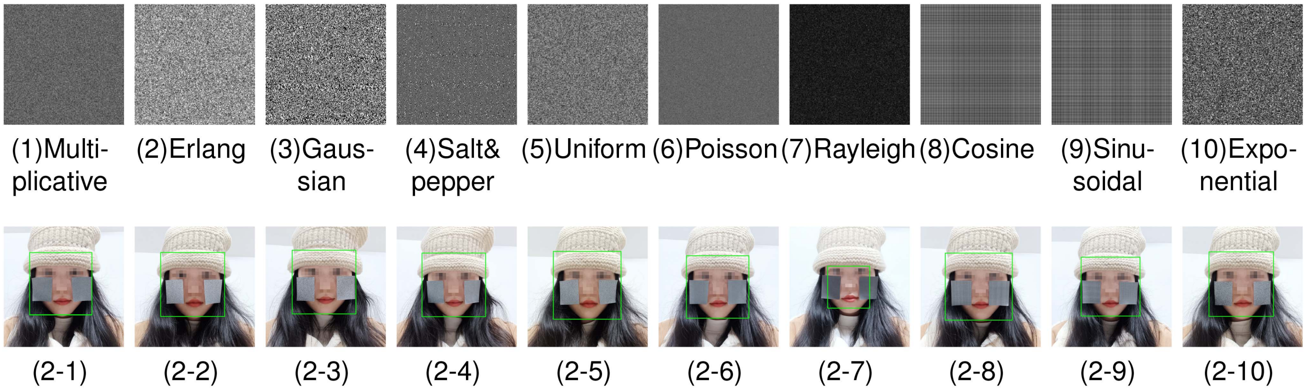
Fig. 10 shows our results. In the figure, the notations L30, L60, L90 and D30, D60, D90 respectively denote experimental measurements under two illumination conditions and three sensor-object distances, defined as:

- L (Light condition): illumination intensity $\geq$ 100 lx
- D (Dark condition): illumination intensity $\leq$ 10 lx
- 30/60/90: distance between the imaging sensor and the target object (unit: cm)

We evaluated each group's confidence levels under adversarial attacks (orange) versus non-adversarial conditions (green), with a number in each bar representing the confidence reduction caused by EoA adversarial attacks. A high confidence level

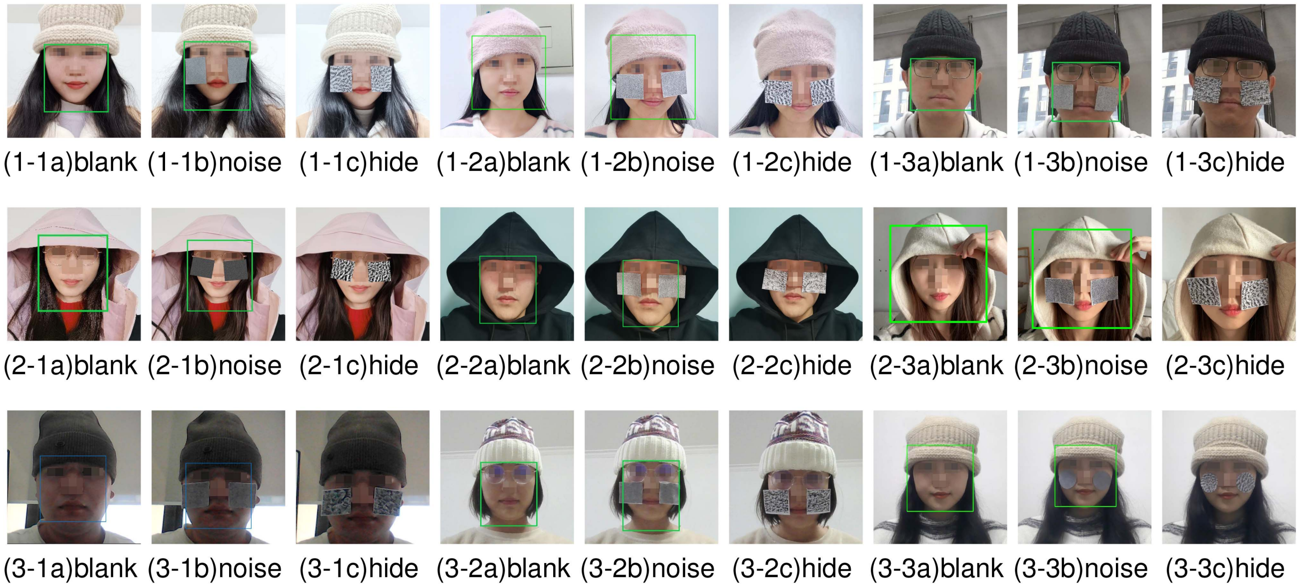
[6] https://github.com/bubbliiiing/retinaface-pytorch
[7] https://github.com/Star-Clouds/CenterFace/tree/master/prj-python
[8] https://github.com/hpc203/yolov8-face-landmarks-opencv-dnn
[9] https://github.com/deepinsight/insightface/tree/master/detection/scrfd
[10] https://github.com/dlunion/DBFace/tree/master
[11] https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.


(1)Multiplicative (2)Erlang (3)Gaussian (4)Salt& pepper (5)Uniform (6)Poisson (7)Rayleigh (8)Cosine (9)Sinusoidal (10)Exponential

(2-1) (2-2) (2-3) (2-4) (2-5) (2-6) (2-7) (2-8) (2-9) (2-10)

Fig. 12. Random patches (row 1) and the corresponding detection results when wearing them (row 2).


(1-1a)blank (1-1b)noise (1-1c)hide (1-2a)blank (1-2b)noise (1-2c)hide (1-3a)blank (1-3b)noise (1-3c)hide

(2-1a)blank (2-1b)noise (2-1c)hide (2-2a)blank (2-2b)noise (2-2c)hide (2-3a)blank (2-3b)noise (2-3c)hide

(3-1a)blank (3-1b)noise (3-1c)hide (3-2a)blank (3-2b)noise (3-2c)hide (3-3a)blank (3-3b)noise (3-3c)hide

Fig. 13. Representative hat-wearing experiments (blank: No patch, noise: Random patch, and hide: No face detected).

indicates a higher likelihood that a face has been detected. Among all models, *RetinaFace* exhibited the highest face detection accuracy without patches across different distances and lighting conditions, while also achieving significantly effective evasion results when using EoA-generated patches.

The FD evasion experiment on *UltraFace* in light environments yielded strong results; however, due to its inability to detect faces in dark conditions, its evasion effectiveness in low-light settings could not be assessed. Overall, analysis of other models indicates that EoA effectively enables face detection evasion across most environments and models. Notably, while *YOLO8Face* and *DBFace* showed some resistance to patches in light environments, they failed to defend against the evasion attack in dark conditions.

We observed that when other conditions are the same, the FD confidence levels are generally higher in light condition than in dark condition, no matter with or without patches. The low confidence level with a patch in the dark environment indicates easier evasion from FD systems.

As for the distance, we observed that distance has a varying impact on confidence levels. For example, under light condition, for the *RetinaFace*, *DBFace*, and *UltraFace* model, the patch makes confidence level decline when distance increase from 30 to 60, but rise again when distance increase from 60 to 90. For the other three models, the patch makes the confidence level increase as the distance increases, meaning it's easier to detect faces when distance increases. Analyzing the model parameters and input images, we found that greater distances could possibly result in reduced clarity of the adversarial patch captured by the camera. Additionally, some models significantly resize input images, further degrading patch clarity. As the patch becomes less distinct, its ability to divert the model's attention weakens, reducing the likelihood of evasion.

E. Ablation Study

In EoA, we construct an ensemble of M publicly available models to compute PAHM. In this section, we conduct ablation

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

TABLE III
CONFIDENCE LEVEL OF DIFFERENT SIZES OF PATCHES (LOWER CONFIDENCE INDICATES BETTER EVASION)

Patch Width	FD model		
	RetinaFace	CenterFace	Yolo8Face
3cm	93.1%	86.9%	76.2%
4cm	67.1%	61.9%	55.1%
5cm	**32.4%**	**52.4%**	**41.9%**

TABLE IV
CONFIDENCE LEVEL OF DIFFERENT POSITIONS OF PATCHES (LOWER CONFIDENCE INDICATES BETTER EVASION)

Position	FD model		
	RetinaFace	CenterFace	Yolo8Face
Cheeks	**32.3%**	**40.7%**	**55.6%**
Cheekbones	59.2%	46.5%	60.8%
Corners of the mouth	73.2%	87.7%	57.5%

TABLE V
COMPARISON OF CONFIDENCE LEVELS BETWEEN COLORED AND BLACK-AND-WHITE PATCHES (LOWER CONFIDENCE INDICATES BETTER EVASION)

Color	Environment		
	L30	L60	L90
Colored	57.3%	52.4%	74.3%
Black-and-white	32.4%	21.4%	31.2%
	D30	D60	D90
Colored	31.4%	47.1%	69.1%
Black-and-white	19.1%	7.6%	11.2%

experiments to illustrate the influence of the parameter M on the effectiveness of the proposed attack. We use the same experimental settings and systems as detailed in Section IV-C, and vary the number of public models utilized for generating adversarial patches: (1) $M = 1$ (MTCNN); (2) $M = 2$ (MTCNN + Pyramidbox); (3) $M = 3$ (MTCNN + Pyramidbox + Facebox). The corresponding results of the attacks are presented in Table II.

The results show that increasing the value of M (representing the number of models in the ensemble) results in a higher attack success rate. This demonstrates that using three public models (i.e., $M = 3$) is better than using just one or two (i.e., $M = 1$ or 2) models. Specifically, using the 60 cm distance as the example, the efficacy improvement from raising M from 1 to 2 is 40%, and the gain from raising M from 2 to 3 amounts to 12%. This pattern reveals that as the number of models included increases, the incremental gain begins to diminish. Considering that each increment of M incurs additional computational overhead, the setting of $M = 3$ in our implementation strikes a balance of delivering satisfactory performance and effectiveness while avoiding excessive computational burdens. Further experiments, which involve random model selection, confirm that once the value of M is determined, altering the combinations of FD models does not substantially impact efficacy. This finding provides a solid foundation for model selection as detailed in Section IV-A.

F. Impact of Patch Configurations

In previous sections, we used a patch with a size of 5 cm × 6 cm to carry out the experiment. Larger patches, which cover more of the face, result in a more pronounced adversarial effect as the patch area increases. To more comprehensively explore the potential of this approach in practical applications, this section investigates the evasion effect of smaller patches in the physical world. Specifically, we conducted additional tests using patches with widths of 3 cm and 4 cm, and carefully evaluated their performance under three face detection models: *RetinaFace*, *YOLO8Face*, and *CenterFace*. These results were then compared with those from 5cm-wide patches. To ensure fairness during training, all initial patches were placed symmetrically on both sides of the nasal wings, without covering any key facial features. The results are shown in the Table III, where lower confidence level indicates better evasion results.

The experimental results align with our expectations: as the size of the adversarial patch decreases, its effectiveness drops rapidly across all three models. This indicates that the 5 cm-wide patch represents the smallest area that still retains strong adversarial properties. In future work, we aim to further explore the potential for reducing patch size without compromising effectiveness.

Next, we examined the impact of patch position on the face. Using the position from previous experiments as a baseline, we compared it with alternative placements, shifted outward and downward, while ensuring no key facial features were obstructed. We conducted tests on the *RetinaFace*, *CenterFace*, and *YOLO8Face* models, capturing 20 images under *L30*, *L60*, and *L90* lighting conditions, and calculated the average confidence score as the evaluation metric. The results, summarized in Table IV, indicate that patches applied to the cheeks (closer to the eyes than the mouth) yield the most effective results. In general, placements near the eyes proved more effective than those near the mouth.

Finally, we conducted an experimental analysis on the color of the patch. Using black-and-white patches as a baseline, we tested the effect of colored patterns on the adversarial patches. The training process kept the same parameters and input images, changing the initial mask from pure black to a random color, thus generating colored adversarial patches. We evaluated these patches on existing FD models, with the attack effects of color patch versus black-and-white patch on *RetinaFace* presented in Table V. The results show that colored patches perform less effectively than black-and-white patches under both bright and dark conditions. Our analysis suggests that, during the process of generating color patches, some adversarial characteristics are encoded in the color itself, which can be affected by factors like lighting intensity and camera hardware in real-world settings. This leads to discrepancies between the patch colors captured by the FD system and the original colors, diminishing the adversarial effect, particularly in dark conditions. Additionally, we found that training color patches requires three channels, whereas black-and-white patches only need one channel, making the training of color patches more time and resource-intensive. The conditions *L30, L60, L90, D30, D60*, and *D90* are defined in

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

Section IV-D. The confidence levels for *D30*, *D60*, and *D90* are lower than those for *L30*, *L60*, and *L90*, as the face detection (FD) model performs less effectively in dark environments.

G. Video-Based Adversarial Effectiveness Analysis

We’ve created four demonstration videos (available through the same GitHub repository) showcasing our method. One video features no adversarial patch, while the other three include adversarial patches.

When a face is detected, the FD model displays a blue rectangle around it; otherwise, no output is shown. All videos were recorded using the same FD model under consistent lighting conditions and identical patch configurations, as detailed in Sections IV-F and IV-D. To evaluate the robustness of the adversarial patches generated by the EoA algorithm, we recorded videos while introducing variations in head orientation, camera distance, and facial expressions like in a dynamic, real-world environment. The confidence threshold of the FD model was set to match that of a typical commercial face detection system.

We’ve summarized the statistical results from the videos, showing the proportion of frames in which faces were detected versus not detected. 1) In the video featuring adversarial attack at a distance of 30 cm from the camera, there are *575* frames in total: faces were detected in only *3* frames, while the remaining *572* showed no detection; 2) In the video featuring adversarial attack at a distance of 60 cm from the camera, there are *645* frames in total: faces were detected in only *11* frames, while the remaining *634* showed no detection; 3)In the video featuring adversarial attack at a distance of 90 cm from the camera, there are *352* frames in total: faces were detected in only *19* frames, while the remaining *333* showed no detection; 4) In the video without adversarial attacks, recorded at distances of 30, 60, and 90 cm under consistent lighting conditions, all *647* frames resulted in successful face detection.

We analyze video frames by grouping them according to facial features and angles. Our findings show that it’s easier to evade the face detection with patches when the face turns to wider angles than when the face is oriented directly toward the camera. This can be attributed to the FD model’s stronger performance in detecting frontal faces under normal conditions without patches. Overall, our adversarial patch demonstrates great robustness to facial deflection, leading to good evasion results when applied to non-frontal faces.

In addition, we compared the evasion performance across different facial expressions and found that most expressions do not alter the patch’s position, and thus have minimal impact on its effectiveness. Interestingly, certain uncommon facial expressions may even lower the detection accuracy of the FD model.

H. Generating More Imperceptible Patches

In the aforementioned experiments, we manage to fool various FD applications, models, and camera devices, causing them incapable of detecting faces correctly. The adversarial patches we create, despite of their capability to divert the detector’s attention away from the actual facial area, may draw human attention when they are worn on individual’s cheeks. In order to craft adversarial patches that maintain a visually natural appearance (for human imperceptibility), we introduce a novel loss term that takes into account human attention. This approach is inspired by a method described in [83].

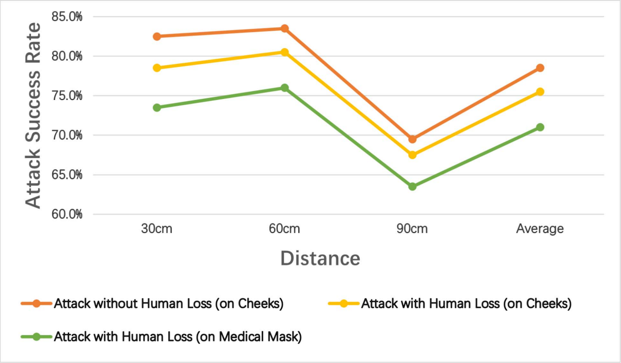


Fig. 14. Case study: The attack success rates against yoloface model.

In details, we create a target pattern, denoted as T_0, which contains a strong semantic connection with the context of the scenario. The human attention loss, represented as L_h, is designed within the following equation to make adversarial patches closely align with this target pattern:

$$\mathcal{L}_h = \|(\theta \cdot (\mathbf{1} - \mathbf{T_0}) + \mathbf{1}) \odot (\mathbf{T}_{adv} - \mathbf{T}_0)\|_2^2 \qquad (6)$$

where θ is the hyper-parameter, $\theta \cdot (\mathbf{1} - \mathbf{T_0}) + \mathbf{1}$ is the weight tensor with the same dimension as $\mathbf{T}_0$, and $\odot$ denotes the element-wise multiplication. Accordingly, (5) can be updated as follows by adding the human attention loss L_h into the original total loss L, where β is a hyper-parameter like γ for balancing.

$$\mathcal{L} = L_{\text{PAHM}} + \gamma L_{TV} + \beta L_h \qquad (7)$$

Case Study: We present a case study to illustrate the practical application and effectiveness of our imperceptible adversarial patches. In response to the global COVID-19 pandemic, transportation hubs have upgraded their FD systems to accommodate the identification of passengers wearing medical masks. In light of this development, we craft adversarial patches to be applied to medical masks, effectively deceiving these advanced FD systems. This aligns closely with the trend of many medical masks featuring visually appealing patterns to attract consumers. Specifically, we select a target pattern (Fig. 11(a)), and then follow (7) to craft adversarial patches based on this pattern (Fig. 11(b)). This is more imperceptible than the patches generated previously. Subsequently we attach this patch to a medical mask (Fig. 11(c)), which can be employed to successfully evade the face detection.

To quantitatively assess the effectiveness of the human attention loss L_h, we select the Yoloface model as a representative victim model, and compare the attack success rates in different scenarios: 1) adversarial mask with L_h; 2) adversarial patch on the cheeks without L_h; 3) adversarial patch on the cheeks with L_h. Fig. 14 presents the comparison results. We notice that the inclusion of the human attention loss term L_h results in only a minor decrease in the success rate. This is a promising outcome

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

as it suggests that the patch becomes more imperceptible while preserving a high level of effectiveness. This also enables the possibility of integrating the patches seamlessly into facial or head coverings to make the patches less noticeable.

V. Limitations, Discussion, Future Works

Limitation in attacking 3-Dimension FD systems: Evaluations in Section IV demonstrate that our attack is capable of compromising common black-box FD apps and models. Nevertheless, it is essential to acknowledge the potential resilience of more advanced systems to our approach. A prime example is Apple's implementation of structured light technology for the Face ID module on the iPhone, which projects small infrared dots onto a user's face and estimates its 3D shape [84]. This advanced technique effectively mitigates FD evasion attempts (including ours) based on 2D images. Our attempts to utilize adversarial patches to subvert the iPhone unlocking function yielded no success; however, we were able to successfully subvert the unlocking functions of the other three smartphones, as they rely on a purely camera-based solution for facial detection. Another example is Lidar-based FD [85], [86], which builds point cloud models to recognize faces. Our attack primarily focuses on targeting the facial surfaces, and hence could not circumvent 3D detection mechanisms. Such advanced detection methods have not been widely commercialized due to the significant expenses associated with sensing. In the future, we plan to enhance our methods to cover these advanced FD methodologies.

Limitation in attacking face recognition systems: In this paper, we focus on attacking the FD systems, which is tasked with identifying the presence of human faces. There exists another category of applications, namely face recognition, which further serves the purpose of recognizing the identity of the face [4], [5]. Numerous works have studied strategies for launching attacks towards the latter category, in both digital [87], [88] and physical domains [89], [90]. Methods have been explored in both white-box [90], [91] and black-box settings [13]. However, these solutions are still distance away from being generally effective when it comes to targeting more diverse face recognition applications in the real world. This sheds light for our future endeavors.

We would like to clarify the difference between face detection evasion attack and the above face recognition adversarial attacks. Face detection (FD) is a fundamental computer vision technique that identifies the presence of human faces in images or videos. It plays a critical role in various applications such as face alignment, face recognition, facial expression analysis, and face tracking. In face recognition adversarial attacks, the attacker aims to cause misclassification of adversarial face images by recognition systems. That is, their objective is prevent the system from correctly matching the target individual with their true identity in the database, or cause the system to mistakenly recognize the target individual as another specific person. In contrast, in face detection evasion, the attacker's goal is to prevent face detection systems from detecting faces in images or to reduce the system's confidence in detected faces, rendering the target individual undetectable.

Extended attacks: We mainly target at face hiding through attacks in this paper. Another type of attacks is appearing attack, which goal is to deceive the detector into recognizing non-face objects, overlaid with patches, as actual faces. Similar concepts have been explored in prior studies for general object detection tasks [34], [71]. In our future work, we intend to extend our EoA approach to incorporate this type of attack. Rather than diverting the detector's attention, the objective can be to direct its focus on adversarial patches, leading to potential wrongful face recognition.

Patch configuration and optimization: Our approach aligns with prior research such as [29] in terms of the dimensions, positioning, and configurations of adversarial patches. This paper, grounded in their empirical parameter settings, prioritizes the investigation of strategies to achieve generality of the solution and maximize attack success rates. Future work can include the systematic exploration of optimal attributes (such as placement, size, and shape) of adversarial patches for attack effectiveness. Considerations may also involve techniques for rendering patches imperceptible to human detection, such as reducing patch size and integrating them seamlessly into facial or head coverings, such as QR code on a hat.

More extreme environmental conditions: When designing adversarial patches, we take into account various environmental factors, such as distance, viewing angles, and lighting conditions. Consequently, our attack remains robust against the specified range of variations. However, more extreme environmental conditions, such as poor lighting, long distances, image blur or distortion, could potentially render our patches ineffective. To further increase the robustness of adversarial patches, in future research we will seek to augment carrier images with effects derived from these extreme conditions.

Possible defense strategies: Currently, there is no well-known FD system specifically designed to blur or eliminate facial patches. Our work targets this vulnerability and demonstrates effective results, highlighting the need for FD systems to be enhanced against interference from such adversarial patches, an insight we aim to contribute to the field. Furthermore, our method serves as a tool to evaluate the security and robustness of face detection models deployed in real-world. When our adversarial patches successfully bypass a target system, it underscores the need to strengthen its defense mechanisms. Based on our experimental experience, we propose the following defense strategies to mitigate the risk of physical-domain adversarial patch attacks in face detection systems:

1) Defense through Image Pre-processing: The generation of adversarial patches depends on input images, where the quality is intricately tied to the camera resolution and shooting distance. For the attack to be effective, the adversarial patch must be fully present in the input image. Missing or damaged parts, as well as variations in camera resolution, can significantly reduce its effectiveness. By introducing frequency domain interference or spatial domain degradation processing in the image pre-processing process between the image acquisition module and the detection module, the effectiveness of adversarial patch attacks can be effectively weakened.

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

2) Generative Defense Framework: The first step is to localize the abnormal patch region on the face. Once identified, a pre-trained diffusion model is used to progressively reconstruct the affected area. This reconstruction is guided by the constraint of preserving the individual's original identity features. Through multi-step iterative sampling, the model gradually restores natural facial textures, effectively transforming the adversarial perturbations into visually plausible and non-aggressive pixels.
3) Enhanced Detection System: Defenders can address the limitations of traditional five-point detection models by: a) enhancing spatial representation with more keypoint detections (e.g., >20 points); b) adopting 3D face detection models to improve face recognition using depth information; c) integrating data from visible light, near-infrared, or thermal imaging to construct a multi-modal face detection system based on diverse physical characteristics; and (d) utilizing images with adversarial patches to fine-tune the original face detection network, thereby increasing robustness against such attacks.

Alternative FD evasion methods: Our review noted several studies that attempt to evade or bypass face detection (FD) using optical interference techniques. These include: [92] generating adjustable, invisible laser perturbations directed at the camera's CMOS sensor; [93] projecting digital adversarial patterns onto the attacker's face using a projector; [32] emitting invisible infrared light spots from a hat-mounted LED; and [33] using near-infrared signals emitted through devices resembling glasses to interfere with the camera. Notably, these methods rely on specialized optical equipments, which may inadvertently affect human vision due to strong light exposure. Although they pursue a similar objective, these methods are not directly comparable to ours.

VI. Conclusion

In this paper, we introduce a physical-world adversarial attack targeting black-box face detection (FD) systems. Leveraging the attention *heat map's* capacity to reflect the FD model's focus, we devise an attention-based loss function to divert the detector's focus, rendering it incapable of identifying faces wearing adversarial patches. The creation the Public Attention *Heat map* (PAHM) is key to enabling the generality and robustness of our adversarial patches, allowing us to successfully compromise black-box FD systems across multiple platforms (smartphones, laptop, suveillance camera, etc.), all without prior knowledge of the detection algorithms. Moreover, we employ a human attention loss function to demonstrate a successful strategy of generating and camouflaging adversarial patches as part of face covering (such as medical masks in a case study), to make our attack more imperceptible while preserving effectiveness in evading FD models.

Acknowledgment

Thanks go to Jingcheng Ju, Haowen Xu, and Zhenlu Tan, for their meaningful contributions to the revisions.

References

[1] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 146–155.

[2] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 109–122.

[3] Y. Tai et al., "Towards highly accurate and stable face alignment for high-resolution videos," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8893–8900.

[4] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1891–1898.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1701–1708.

[6] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.

[7] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 787–796.

[8] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[10] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," in *Proc. 9th Int. Conf. Articulated Motion Deformable Objects*, 2016, pp. 175–184.

[11] N. M. Hussien et al., "Smart system for detecting the entry of authority people in the security facilities based IoT using SURF recognition and Viola-Jones algorithms," *J. Phys., Conf. Ser.*, vol. 1963, 2021, Art. no. 012075.

[12] F. Z. Zhou, G. C. Wan, Y. K. Kuang, and M. S. Tong, "An efficient face recognition algorithm based on deep learning for unmanned supermarket," in *Proc. Prog. Electromagnetics Res. Symp.*, 2018, pp. 715–718.

[13] Y. Dong et al., "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7706–7714.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–11.

[15] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–10.

[16] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 284–293.

[17] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.

[18] P. -Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. -J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, 2017, pp. 15–26.

[19] Q. Huang, I. Katsman, Z. Gu,H. He, S. Belongie, and S. -N. Lim, "Enhancing adversarial example transferability with an intermediate level attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4732–4741.

[20] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2137–2146.

[21] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–22.

[22] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.

[23] R. R. Mekala, A. Porter, and M. Lindvall, "Metamorphic filtering of black-box adversarial attacks on multi-network face recognition models," in *Proc. IEEE/ACM 42nd Int. Conf. Softw. Eng. Workshops*, 2020, pp. 410–417.

[24] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

[25] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[26] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–41, 2020.

[27] J. Sun et al., "Stealthy and efficient adversarial attacks against deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5883–5891.

[28] A. J. Bose and P. Aarabi, "Adversarial attacks on face detectors using neural net based constrained optimization," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process.*, 2018, pp. 1–6.

[29] E. Kaziakhmedov, K. Kireev, G. Melnikov, M. Pautov, and A. Petiushko, "Real-world attack on MTCNN face detection system," in *Proc. Int. Multi-Conf. Eng., Comput. Inf. Sci.*, 2019, pp. 0422–0427.

[30] Y. Li, X. Yang, B. Wu, and S. Lyu, "Hiding faces in plain sight: Disrupting AI face synthesis with adversarial perturbations," 2019, *arXiv:1906.09288*.

[31] C. Zhou, H. Jing, X. He, L. Wang, K. Chen, and D. Ma, "Disappeared face: A physical adversarial attack method on black-box face detection models," in *Proc. 23rd Int. Conf. Inf. Commun. Secur.*, 2021, pp. 119–135.

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

[33] M. Shen, Z. Liao, L. Zhu, K. Xu, and X. Du, "VLA: A practical visible light-based attack on face recognition systems in physical world," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–19, 2019.

[34] D. Song et al., "Physical adversarial examples for object detectors," in *Proc. 12th USENIX Workshop Offensive Technol.*, 2018, pp. 1–10.

[35] N. Guetta, A. Shabtai, I. Singh, S. Momiyama, and Y. Elovici, "Dodging attack using carefully crafted natural makeup," 2021, *arXiv:2109.06467*.

[36] T. Yamada, S. Gohshi, and I. Echizen, "Privacy visor: Method based on light absorbing and reflecting properties for preventing face image detection," in *Proc. 2013 IEEE Int. Conf. Syst., Man, Cybern.*, 2013, pp. 1572–1577.

[37] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[38] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1522–1530.

[39] Y. Li, B. Sun, T. Wu, and Y. Wang, "Face detection with end-to-end integration of a convnet and a 3D model," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 420–436.

[40] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis. Image Understanding*, vol. 138, pp. 1–24, 2015.

[41] S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 211–223, Feb. 2016.

[42] S. C. Brubaker, J. Wu, J. Sun, M. D. Mullin, and J. M. Rehg, "On the design of cascades of boosted ensembles for face detection," *Int. J. Comput. Vis.*, vol. 77, pp. 65–86, 2008.

[43] Y. Zhang, X. Xu, and X. Liu, "Robust and high performance face detector," 2019, *arXiv:1901.02350*.

[44] L. Bourdev and J. Brandt, "Robust object detection via soft cascade," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 236–243.

[45] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 122–138.

[46] H. Qin, J. Yan, X. Li, and X. Hu, "Joint training of cascaded CNN for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3456–3465.

[47] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image Vis. Comput.*, vol. 32, pp. 790–799, 2014.

[48] S. S. Farfade, M. J. Saberian, and L. -J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 643–650.

[49] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5325–5334.

[50] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 82–90.

[51] S. Yang, P. Luo, C. -C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3676–3684.

[52] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Deep Learning for Biometrics*. Cham, Switzerland: Springer, 2017, pp. 57–79.

[53] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[54] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 650–657.

[55] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[56] J. Wang, Y. Yuan, and G. Yu, "Face attention network: An effective face detector for the occluded faces," 2017, *arXiv:1711.07246*.

[57] C. Zhang, X. Xu, and D. Tu, "Face detection using improved faster RCNN," 2018, *arXiv:1802.02142*.

[58] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 91–99.

[59] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4885–4894.

[60] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 797–813.

[61] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, "Face detection through scale-friendly deep convolutional networks," 2017, *arXiv:1706.02863*.

[62] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^{3}FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 192–201.

[63] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Consistent optimization for single-shot object detection," 2019, *arXiv:1901.06563*.

[64] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849.

[65] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[66] W. Feng, B. Wu, T. Zhang, Y. Zhang, and Y. Zhang, "Meta-attack: Class-agnostic and model-agnostic physical adversarial attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7767–7776.

[67] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[68] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 284–293.

[69] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector," in *Mach. Learn. Knowl. Discov. Databases: Eur. Conf.*, 2018, pp. 52–68.

[70] S. Thys, W. V. Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 49–55.

[71] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2019, pp. 1989–2004.

[72] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, "Universal adversarial attack on attention and the resulting dataset damagenet," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2188–2197, Apr. 2022.

[73] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. -R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Cham, Switzerland: Springer, 2019.

[74] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5202–5211.

[75] Y. Xu, W. Yan, G. Yang, J. Luo, T. Li, and J. He, "CenterFace: Joint face detection and alignment using face as point," *Sci. Program.*, vol. 2020, 2020, Art. no. 7845384.

[76] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–9.

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.

[77] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[78] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "FaceBoxes: A CPU real-time face detector with high accuracy," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2017, pp. 1–9.

[79] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D, Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.

[80] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9185–9193.

[81] J. Li et al., "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5055–5064.

[82] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[83] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8561–8570.

[84] Apple's new face id system uses a sensing strategy that dates back decades. [Online]. Available: https://www.popsci.com/apple-face-ID/

[85] X. Li, J. Wan, Y. Jin, A. Liu, G. Guo, and S. Z. Li, "3DPC-Net: 3D point cloud network for face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2020, pp. 1–8.

[86] Z. Zhang, F. Da, and Y. Yu, "Data-free point cloud network for 3D face recognition," 2019, *arXiv:1911.04731*.

[87] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2020, pp. 1–10.

[88] L. Yang, Q. Song, and Y. Wu, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *Multimedia Tools Appl.*, vol. 80, no. 1, pp. 855–875, 2021.

[89] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on arcface face ID system," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 819–826.

[90] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Trans. Privacy Secur.*, 2019, pp. 1–30.

[91] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1528–1540.

[92] Y. Wang, Z. Liu, B. Luo, R. Hui, and F. Li, "The invisible polyjuice potion: An effective physical adversarial attack against face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2024, pp. 3346–3360.

[93] D. -L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial light projection attacks on face recognition systems: A feasibility study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3548–3556.

Duohe Ma (Member, IEEE) received the BE and MS degrees from the Harbin Institute of Technology, in 2004 and 2006, respectively, and the Ph.D. degree in information security from the University of Chinese Academy of Sciences, in 2015. He is currently an associate professor with the State Key Laboratory of Information Security, Institute of Information Engineering CAS, and University of Chinese Academy of Sciences. His research interests include AI Security, moving target defense, data security, and cloud security.

Junye Jiang is a doctoral student at the Institute of Information Engineering, University of Chinese Academy of Sciences. He received the bachelor's degree from Northeastern University, China, in 2022. His research interests include deep learning, mobile device security, and AI security, with a particular focus on natural language processing and AI-related threats and defenses.

Xiaoyan Sun (Member, IEEE) received the PhD degree from Penn State University, in 2016. She is currently an associate professor of computer science with Worcester Polytechnic Institute. Her research interests include cybersecurity and digital forensics. She was the publicity co-chair for ACM CCS'20, the TPC and a reviewer for top cybersecurity journals and conferences, such as IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Information Forensics and Security, IEEE/ACM Transactions on Networking, and ACSAC. She is also the vice president for Silicon Valley Cybersecurity Institute, a non-profit organization that promotes cybersecurity research and education.

Zhimin Tang received the bachelor's degree in electronic science and technology from the Beijing University of Posts and Telecommunications, in 2021, and the master's degree in cyberspace security from the University of Chinese Academy of Sciences, in 2024. Zhimin's research interests include moving target defense, data security, and ransomware defense.

Zhenchao Zhang received the BE degree from the Ocean University of China, China, in 2021, and the MS degree from University of Chinese Academy of Sciences, China, in 2024. His research interests include computer vision, artificial intelligence, and AI security. He is particularly interested in AI security, data security, and privacy security.

Kai Chen (Member, IEEE) received the BE and MS degrees from the Ocean University of China, China, in 2009 and 2012, respectively, and the PhD degree in cyberspace security from the University of Chinese Academy of Sciences, China, in 2024. He is currently an assistant professor with the Institute of Information Engineering. His research interests include AI security, moving target defense, and data security.

Jun Dai (Member, IEEE) received the PhD degree from Penn State University, in 2014. He is currently an associate professor of computer science with Worcester Polytechnic Institute. He has authored or coauthored in top-tier academic venues, including NDSS, ACM SIGMOD, IEEE Transactions on Information Forensics and Security, ACM SIGCSE, and IFIP WISE. His research interests include the intersections of networks, distributed systems, artificial intelligence (AI), and cybersecurity, with a recent emphasis on intrusion detection, vulnerability analysis, secure programming, and cybersecurity education.

Authorized licensed use limited to: Worcester Polytechnic Institute. Downloaded on December 07,2025 at 00:11:20 UTC from IEEE Xplore. Restrictions apply.