Adversarial Mask: Real-World Adversarial Attack Against Face Recognition Models

Alon Zolfi[1], Shai Avidan[2], Yuval Elovici[1], Asaf Shabtai[1]
[1] Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev
[2] School of Electrical Engineering, Tel Aviv University
zolfi@post.bgu.ac.il, avidan@tauex.tau.ac.il, {elovici,shabtaia}@bgu.ac.il

Abstract

Deep learning-based facial recognition (FR) models have demonstrated state-of-the-art performance in the past few years, even when wearing protective medical face masks became commonplace during the COVID-19 pandemic. Given the outstanding performance of these models, the machine learning research community has shown increasing interest in challenging their robustness. Initially, researchers presented adversarial attacks in the digital domain, and later the attacks were transferred to the physical domain. However, in many cases, attacks in the physical domain are conspicuous, requiring, for example, the placement of a sticker on the face, and thus may raise suspicion in real-world environments (e.g., airports). In this paper, we propose Adversarial Mask, a physical adversarial universal perturbation (UAP) against state-of-the-art FR models that is applied on face masks in the form of a carefully crafted pattern. In our experiments, we examined the transferability of our adversarial mask to a wide range of FR model architectures and datasets. In addition, we validated our adversarial mask effectiveness in real-world experiments by printing the adversarial pattern on a fabric medical face mask, causing the FR system to identify only 3.34% of the participants wearing the mask (compared to a minimum of 83.34% with other evaluated masks).

1. Introduction

For the past two years, the coronavirus has impacted every aspect of our lives, and its impact will continue for the foreseeable future. Since its emergence, various suggestions have been made to reduce its spread. While the effectiveness of some actions is questionable, there is no doubt that face masks are a key factor in preventing the spread of the virus in crowded and enclosed spaces.

The widespread adoption of face masks and the ever-increasing use of deep learning-based FR models in everyday systems can be leveraged to perpetrate targeted adversarial attacks that will enable attackers to evade such models and compromise their robustness, without raising an alarm.

Figure 1: Illustrating the adversarial pattern printed on a fabric mask, which results in the failure of the FR system to detect the person wearing it, compared to the detection without a mask as well as with a standard disposable mask.

Adversarial attacks in the computer vision domain have gained a lot of interest in recent years, and various ways of fooling image classifiers [33, 10, 17, 32, 24] and object detectors [8, 27, 31, 34, 45] have been proposed. Attacks against FR systems have also been shown to be effective. For example, research has demonstrated that face synthesis in the digital domain can be used to fool FR models [41]. In the physical domain, some methods involved wearing adversarial eyeglasses [28], projecting lights on human faces [30], wearing a hat containing an adversarial sticker [16], and using adversarial makeup [11]. However, the proposed attacks are conspicuous and do not allow the attacker to blend in naturally in real-world scenarios, potentially triggering defense systems.

In this work, we propose a *universal* adversarial attack that can be used to physically evade FR systems; in this case, an adversarial pattern is printed on a fabric face mask, as shown in Figure 1. To create the adversarial pattern, we use a gradient-based optimization process that causes *all* identities wearing the mask to be misclassified by the FR model. We first demonstrate the attack's ability to fool state-of-the-art models (e.g., ArcFace [7]) in the digital domain by applying the face mask to every facial image in the

dataset (dynamically) using 3D face reconstruction. Then, we print the adversarial pattern on an actual fabric face mask and test it under real-world conditions. The results in the digital domain show that our adversarial mask performs better than all evaluated masks and is transferable to other models. In the physical domain, we show that 96.66% of the participants wearing our mask were able to evade the detection of the system.

The contributions of our research can be summarized as follows:

- We are the first to present a universal adversarial attack that fools FR models, i.e., we craft a single perturbation that causes the FR model to falsely classify all attackers (those enrolled in the system) as unknown identities, even under diverse conditions (angles, scales, etc.)
- In the digital domain, we study the transferability of our attack across different model architectures and datasets.
- We present a novel digital masking method that can accurately place any kind of mask on any face, regardless of the position of the head. This method can be used for other computer-vision tasks (e.g., training masked-face detection models).
- We present an inconspicuous practical (physical) attack printed on a fabric face mask that can be used in real-world scenarios without raising an alarm, due to the covid-19 pandemic.
- We propose various countermeasures that can be used during the FR model training and inference phases.

2. Background & Related Work

In this section, we present previous studies that are relevant to this paper. In addition, we provide background knowledge about the main components of FR systems.

2.1. Adversarial Attacks

Adversarial attacks against machine learning models have been extensively studied over the past few years and researchers have proposed various ways of fooling deep neural networks (DNNs) in the computer vision domain. These methods can be roughly categorized as: (a) digital attacks, which operate in the digital space, or (b) physical attacks, which operate in the real world.

Digital Attacks. Initially, attacks in the digital domain aimed at fooling classification models were introduced [33, 10, 18, 17, 25, 32, 24]. While those earlier attacks are based on methods that generate a perturbation for a single image, Moosavi-Dezfooli *et al*. [23] proposed *universal* adversarial perturbations (UAP), which enables any image that is blended with the UAP to fool a DNN. Digital attacks on models that perform more complex computer vision tasks (e.g., face recognition and object detection) have also emerged. Liu *et al*. [20] presented a patch-based attack on object detectors called *DPatch*. By placing a digital patch in the corner of an image (i.e., not on the targeted object itself), they were able to deceive Faster R-CNN [27] and YOLO [26]. Yang *et al*. [41] designed a digital patch which is placed on a person's forehead to deceive face detectors. Recent work on FR models was presented by Deb *et al*. [6], who proposed automated adversarial face synthesis, using a generative adversarial network to create minimal perturbations. However, these attacks only raise the potential threat hidden in such models but cannot be transferred to the physical world.

Physical Attacks. In most of the attacks proposed for the physical domain (i.e., real world), the perturbation is crafted in the digital space, just like attacks in the digital domain. However, they differ, because real-world constraints are considered throughout the process of generating the perturbation, and these constraints allow the perturbations to transfer more easily to the physical world.

In recent years, attacks on object detectors have gained attention. Eykholt *et al*. [8] proposed attaching black and white stickers on a stop sign in a certain pattern to fool image classifiers. Chen *et al*. [5] printed stop signs contain adversarial patterns that evaded detection by Faster R-CNN [27], and Sitawarin *et al*. [31] deceived autonomous car systems by crafting toxic traffic signs that look similar to the original traffic signs.

Methods against person detectors have also been proposed. Thys *et al*. [34] suggested a method that successfully evaded YOLOv2 [26], by attaching a small cardboard plate to a person's body. Continuing this line of research, other studies involved printing adversarial patterns on t-shirts, which resulted in a more realistic article of clothing that blends in the environment more naturally [37, 39, 14].

A slightly different approach, in which the perturbation affects the sensor's perception of the object was first introduced by Li *et al*. [19]. The authors applied a translucent patch on the camera's lens to fool image classifiers. Then, Zolfi *et al*. [45] improved this technique to fool object detectors on *all* class instances.

Numerous studies have demonstrated different ways of fooling FR systems. For example, Shen *et al*. [30] introduced the *visible light-based attack*, where lights are projected on human faces. Other studies showed that carefully applied makeup patterns can negatively affect the performance of FR systems [43, 11]. Accessories were also shown to be effective; for example, Sharif *et al*. [28] suggested wearing adversarial eyeglass frames that were crafted using gradient-based methods. Later, GAN methods were used to generate an enhanced version of the adversarial eyeglass frames [29]. Recently, Komkov *et al*. [16] printed an adversarial paper sticker and placed it on a hat to fool the state-of-the-art *ArcFace* [7] FR model. However,

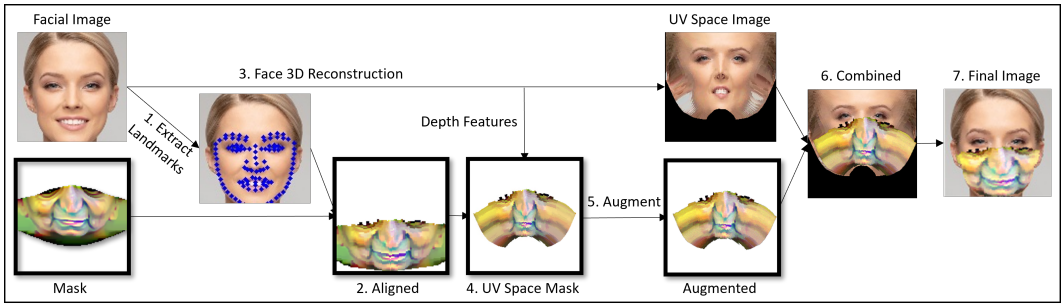


Figure 2: Overview of our mask placement method pipeline.

when implemented on a person, these methods may call attention to the person by causing them to stand out in a crowd given their unnatural appearance. In contrast, we propose a method in which the perturbation is placed on a face mask, a safety measure widely used in the COVID-19 era; in addition, unlike prior work in which the proposed attacks target a single image or person, our universal attack can be applied more widely.

2.2. Face Recognition Systems

The end-to-end procedure of a fully automated facial recognition system consists of several main steps:

- Record - a camera records the environment and then produces a series of frames (a video stream).
- Detect - each frame is analyzed and processed by a face detector to extract cropped faces.
- Preprocess - the cropped faces are aligned according to the FR model's alignment method.
- Embed - the aligned facial images serve as input to an FR model f that maps a facial image I_{face} to a vector $f(I_{face})$, also referred to as an *embedding* vector.
- Compare and verify - the embedding vector is compared to a list of precalculated embedding vectors (also referred as ground-truth embedding vectors) using a similarity measure. The identity with the highest similarity score is marked as a potential candidate; consequently, the candidate identity is only confirmed if its similarity score surpasses a predefined verification threshold (which depends on the system's use case).

Face recognition models can be categorized by two main attributes, the model's backbone, and head, both of which are involved in the training phase. The main architecture used as the backbone in these models is the ResNet [13] architecture, which varies in terms of the number of layers it contains, which is also referred to as the backbone *depth*. On top of the backbone selected, an additional layer (or more) is added, usually containing a novel loss function that is used to train the backbone weights, also referred to as the backbone *head* [7, 35, 22]. Later, when the FR model is used for inference, only the backbone layers are kept to generate the embedding vector.

3. Method

The objective of our research is to generate an adversarial pattern that can be printed on a face mask and cause face recognition systems to misclassify a person's identity. Further, we aim to create an adversarial pattern that is: (a) universal - it must be effective on any identity from multiple views and angles, and at multiple scales, (b) practical - the pattern should remain adversarial when printed on a fabric mask in the real world, and (c) transferable - it must be effective on different models (backbone depths and heads).

3.1. Creating an Adversarial Mask

In this research, we strive to produce an adversarial patch in the form of a face mask. In order to digitally train our adversarial mask, we first need to simulate the patch overlay on a person's face in the real world.

Therefore, we use 3D face reconstruction to digitally apply a mask on a facial image. Feng *et al*. [9] introduced an end-to-end approach called *UV position map* that records the 3D coordinates of a complete facial point cloud using a 2D image. This map records the position information of 3D face and provides dense correspondence to the semantic meaning of each point in the UV space. This method allows us to achieve near-real approximation of the mask on the face, which is essential to the creation of a practical patch.

More formally, we consider our mask $M_{\text{adv}} \in \mathbb{R}^{w \times h \times 3}$ and a rendering function $\mathcal{R}_\theta$. The rendering function (partially inspired from [36]) takes a mask M_{adv} and a facial image X_{face}, and applies the mask on the face.

As shown in Figure 2, the pipeline of the mask's placement is as follows: given a facial image, we first find the landmark points in order to align the mask with the correct location on the original facial image. Then, we input the facial image to the 3D face reconstruction model. The output of the model is used for two purposes: (a) to transfer the original image to the UV space, and (b) to extract the face's depth features to transfer our mask to the UV space. Moreover, to improve the robustness of our patch, we randomly apply location- and color-based transform augmentations:

- Location-based - We add random translation and rotation to simulate possible distortions in the mask placement on the face in the real world.
- Color-based - we add random contrast, brightness, and noise to simulate changes in the appearance of the patch due various possible factors (e.g., lighting, noise or blurring caused when the camera captures the image).

These transformations are parameterized by θ.

Finally, we apply the UV space mask to the UV space facial image and reconstruct the combined image resulting in a masked face image.

Usually, adversarial attacks that employ textile-like objects (e.g., wearable t-shirt [38, 40]), use thin plate splines (TPSs) [2] to simulate fabric distortions. In contrast to these studies, although we aim to craft a textile-based mask, in our case, the mask form on the face remains steady and is not subject to significant distortions. In addition, our 3D approach allows us to simulate smaller distortions (e.g., caused by the nose shape) without actively using TPSs.

Above all, it is important to note that the entire process presented is completely differentiable and allows us to backpropagate and update the mask pixels.

3.2. Patch Optimization

An FR model $f : \mathcal{X}^{w\times h\times 3} \rightarrow \mathbb{R}^N$ receives a face image $x \in \mathcal{X}$ as input and outputs the embedding representation. In our case, the model takes as input a masked face image $\mathcal{R}_\theta(M_{adv}, x)$ and outputs an embedding vector. Since our goal is to make an attacker unknown to FR models, we want to find a patch M_{adv} that will decrease the similarity between the output embedding and the ground-truth embedding e_{gt} (precalculated). Therefore, to minimize the similarity (in our case, the cosine similarity) between the embedding vectors, we use the following loss function:

$$\ell_{sim}(M_{adv}) = \mathbb{E}_{\theta,x}[\cos(f(\mathcal{R}_\theta(M_{adv}, x)), e_{gt})] \quad (1)$$

Since our method is not system-dependent (i.e., using fixed verification threshold determined by a specific use case), we aim to decrease the similarity to the fullest extent possible, in order to perform the most successful attack.

To improve the mask's transferability to other models, we train our patch using an ensemble of FR models. We replace ℓ_{sim} with the following:

$$\ell_{sim}(M_{adv}) = \mathbb{E}_{\theta,x} \sum_j \cos(f^{(j)}(\mathcal{R}_\theta(M_{adv}, x)), e_{gt}^{(j)}) \quad (2)$$

where $f^{(j)}$ denotes the j^{th} model and $e_{gt}^{(j)}$ denotes the embedding representation calculated using the j^{th} model.

We also include the *total variation (TV)* [28] factor to ensure that the optimizer favors smooth color transitions between neighboring pixels and is calculated on the mask pixels as follows:

$$\ell_{TV} = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2} \quad (3)$$

when $\ell_{TV}(M_{adv}) \rightarrow 1$, neighboring pixels are not similar, resulting in a high penalty to the mask.

To be more precise, since ℓ_{sim} output is in the range of $[-1, 1]$ and ℓ_{TV} output is in the range of $[0, 1]$, we transform ℓ_{sim} so it is in the same range ($[0, 1]$); thus, we replace ℓ_{sim} with the following:

$$\ell_{sim}(M_{adv}) = \mathbb{E}_{\theta,x} \sum_j \frac{\cos(f^{(j)}(\mathcal{R}_\theta(M_{adv}, x)), e_{gt}^{(j)}) + 1}{2} \quad (4)$$

Finally, the optimization problem we solve is as follows:

$$\min_{M_{adv}} [\ell_{sim}(M_{adv}) + \lambda * \ell_{TV}(M_{adv})] \quad (5)$$

where λ is set at a small value.

4. Evaluation

In our evaluation, we first run experiments in the digital domain by applying the mask to facial images, using the rendering function R_θ (as explained in Section 3). Then, we evaluate the performance of our adversarial pattern in the physical domain (i.e., real world) by printing it on a fabric mask.

Models. In our evaluation, we used three different types of heads, which are considered state-of-the-art: ArcFace [7], CosFace [35], and MagFace [22]. Specifically, we use pretrained models which were trained using ArcFace and CosFace heads [1], with four different ResNet depths (18, 34, 50, and 100) each, and a pretrained ResNet100 backbone originally trained with the MagFace [22] head, for a total of nine different models. We examine multiple training variations, using one or more (i.e., ensemble) models to train the adversarial mask and then test it on the same training models (i.e., white-box setting) to evaluate the performance. We also evaluate the transferability of our mask to other unknown models (i.e., black-box setting).

Datasets. Throughout this paper, we use three state-of-the-art datasets in the face recognition domain: CASIA-WebFace [42], CelebA [21], and MS-Celeb [12].

For the training phase, we use 100 different identities (50 men and 50 women) from the CASIA-WebFace dataset. We extract five random facial images for each identity, for a total of 500 facial images for training the adversarial mask.

For the evaluation phase, we use 200 identities from each dataset (an equal number of men and women from each dataset), evaluating both the performance on the same distribution (different identities from the CASIA-WebFace

dataset, ∼20k images) and the transferability to other datasets (CelebA and MS-Celeb, ∼6k and ∼24k images, respectively).

Metrics. In our experiments, we quantify the performance of our attack as the ability to degrade the similarity score - specifically the cosine similarity. In the physical domain, we also quantify the effectiveness of our attack using two additional metrics, each of which relates to a different stage of an end-to-end FR system:

- Recognition rate (RR) -

$$RR = \frac{|F_{rec}|}{|F_{det}|} \tag{6}$$

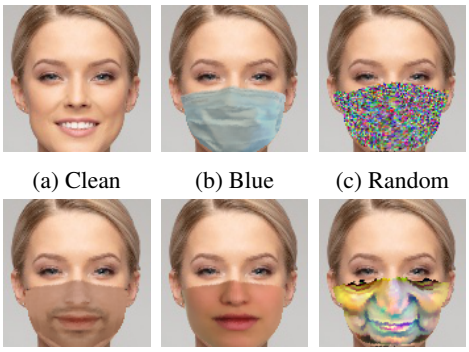
  where $|F_{rec}|$ denotes the total number of frames in which the identity was correctly recognized (the cosine similarity between the ground-truth embedding and the output embedding surpasses the verification threshold), and $|F_{det}|$ denotes the total number of frames in which a face was detected and analyzed by the FR system.
- Persistence detection - since the goal of our adversarial mask is to ensure that an attacker is not identified by the system, we propose a metric that indicates whether the goal was met. An attacker is considered as identified if, within a window of $N_{\text{sliding window}}$ frames, the attacker was recognized in $N_{\text{recognized}}$ frames (where $N_{\text{recognized}} \leq N_{\text{sliding window}}$).

Implementation details. The models we work with in this research only take size $3 \times 112 \times 112$ facial images as input. Therefore, we first resize and align our images to fit these models. We set the size of our patch to be $3 \times 60 \times 112$ to avoid significant downsampling when dynamically rendering the mask to the facial image (using R_θ) due to the small size of the resized facial images, and we set the initial color of the mask to white.

Since our implementation is entirely differentiable, we use an automatic differentiation tool kit (PyTorch) to optimize the mask pixels using the backpropagation algorithm. The pixels are updated using the Adam optimizer [15], where the initial learning rate is set at 10^{-2} and the rate decays if the loss has not improved for two consecutive epochs, until a minimum learning rate of 10^{-6} is reached.

To dynamically detect the landmarks of the original faces, we use a lightweight network [3] that is based on the MobileFaceNet [4] architecture. We chose a small network to keep our end-to-end masking process fast yet dynamic.

Types of face masks evaluated. We compare the effectiveness of our mask with the following masks: (a) Clean - the original facial image without a mask, (b) Adv - our optimized adversarial mask, (c) Random - a mask with randomly colored pixels, (d) Blue - a standard disposable blue mask (simple black and white masks were also tested and yielded the same results). In addition, due to our trained mask's resemblance to a human face, the lower face area of a female and male are used as control masks and will be referred to as *Female Face* and *Male Face*, respectively. The masks compared in our evaluation are shown in Figure 3.



(a) Clean (b) Blue (c) Random

(d) Male Face (e) Female Face (f) Adv

Figure 3: Examples of facial images when various masks are digitally applied to them. (b)-(e) masks placement is partially inspired from [36].

Evaluation setup. Since the state-of-the-art models discussed above were not specifically designed to address the issue of masked faces, we first examine the model's performance on a number of simple face masks. For this evaluation, we use 100 identities from the CASIA-WebFace dataset, where five random images of each identity are used to calculate the ground-truth embedding, and the rest of the images are applied with different types of masks. We then calculate the cosine similarity between the ground-truth embedding and the masked-face embedding images.

To the best of our knowledge, the scientific community has not reached a consensus on the way in which masked face images should be dealt with by FR models. Therefore, we present two approaches of generating the ground-truth embedding: (a) the current approach for unmasked face models - averaging the embedding vectors of the original images only, and (b) an extension to the first approach - in addition to the original images, we create a masked face version for each image (the specific mask is randomly chosen between blue, black, and white) and average the embedding vectors over the two versions of the images.

In Table 1 we can see that although the first approach (w/o Mask) performs better on unmasked images, its performance on masked images is unsatisfactory. On the other hand, the cosine similarity for the second approach (w/Mask) only slightly decreases the cosine similarity on unmasked images (∼0.05 decrease) and performs significantly better on masked images (∼0.1-0.15 increase).

Thus, throughout this section the results we present are obtained using the second approach (the ground-truth embedding used for the training procedure remains the same).

Table 1: Cosine similarity comparison between two ground-truth embedding generation methods on the Resnet100 backbone originally trained with an ArcFace head

Mask Type	Cosine Similarity	
	wo/Mask*	w/Mask**
No mask	**.732**	.682
Blue	.399	**.547**
Black	.407	**.549**
White	.428	**.561**

*Embedding vectors created using original facial images.
**A masked version of the original images is added to the embedding calculation.

It is important to note, that by choosing the second approach, we increase the difficulty of deceiving these models, since the ground-truth embedding vectors encapsulate the use of a mask.

4.1. Results

In this section, we present the results both in the digital and physical domains.

4.1.1 Digital Attacks

We conduct digital experiments to quantify our adversarial mask's effectiveness using the rendering function R_θ (See Section 3), which allows us to dynamically apply masks to the facial images in the test set.

Effectiveness of the adversarial mask in white-box setting. We examine the effectiveness of our attack in a *white-box* setting in which our mask is optimized and tested on the ResNet100 backbone and ArcFace head. As shown in Figure 4, our adversarial mask has more significant impact than the no mask case, in which the average cosine similarity decreased from ~ 0.7 to ~ 0.1. As the case of no mask images represents the upper bound of the cosine similarity, we also performed a targeted attack in which a mask is tailored to each person, to determine the lower bound. The targeted mask results are averaged across all identities in the test set. We can see that the universal mask performs almost as good as a tailor-made mask (~ 0.1 difference). The tailor-made masks propose an attack that is more difficult to detect, since the adversarial pattern varies among different identities. In addition, while the female face and male face control masks are also able to decrease the cosine similarity to a lower level (~ 0.45), our mask outperforms them for almost all tested identities.

Transferability across backbone depth. We examine whether our mask can deceive FR models it was not trained on. Since the majority of the models use the ResNet architecture, we evaluate the performance across different depths of ResNet trained with the ArcFace head. The results are presented in Figure 5a. First of all, we can see that the use of our adversarial mask can cause the cosine similarity to decrease regardless of the model used for training. It can also

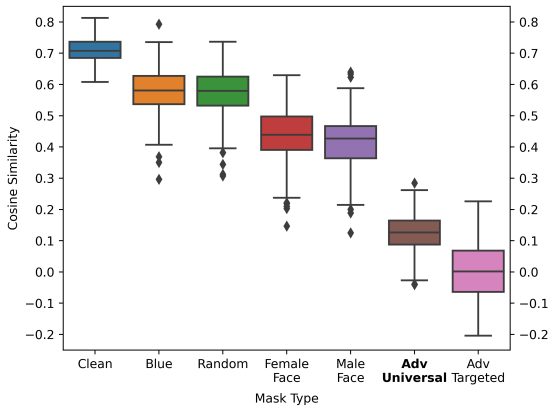


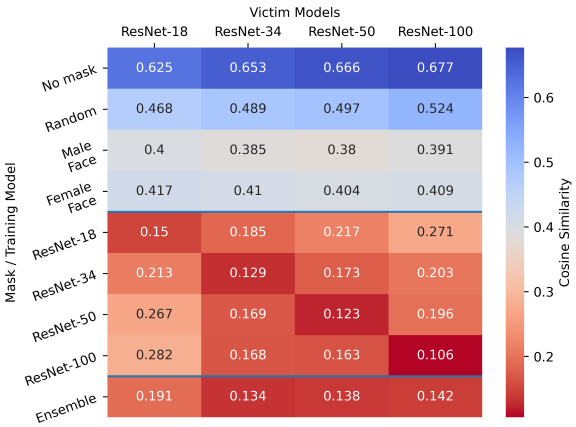
Figure 4: Cosine similarity score across different masks. 'Adv Universal' represents our optimized universal mask, and 'Adv Targeted' represents a tailor-made mask for each identity.

be seen that our attack generalizes better to unknown models whose architecture depth is closer to that of the trained model. For example, an adversarial mask trained on the a model with 100 layers performs better on the models with 34 and 50 layers (decreasing the cosine similarity to 0.168 and 0.163, respectively) than on the 18-layer model (0.282). In addition, we see that the mask trained on an ensemble of all four models does not outperform a mask trained on a single model in white-box setting, however the ensemble's effectiveness is seen overall models combined.

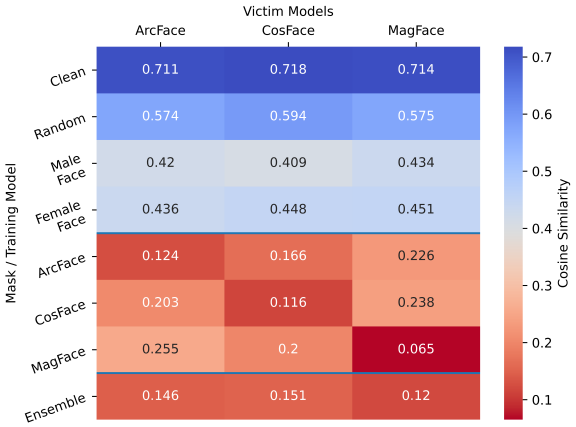
Transferability across network head. We further demonstrate the adversarial mask's transferability across different model heads. We use the ResNet100 backbone in which the weights were learned using one of the following heads: ArcFace, CosFace, and MagFace. As shown in Figure 5b, we observe that our method is head-agnostic, as the decrease in the cosine similarity occurs on all tested models.

However, crafting a mask that was trained using the MagFace model does not generalize as well as the other models where the cosine similarity decreased to 0.065 in the white-box setting but only decreased to 0.255 and 0.2 on the ArcFace and CosFace, respectively. It is interesting to examine the mask learned by each model (presented in Figure 6). Whereas there is a resemblance in the contour of the masks learned, the mask learned using MagFace head(Figure 6c) learns completely different colors than the other two, in some way providing a possible explanation for its lower generalization capability to the other heads. Furthermore, as expected, the ensemble of all three models results in a trade-off - better performance on all three models combined rather than a perfect result on a single model.

Transferability across datasets. We also find our mask to be effective across different datasets. In another experiment, We train our mask using images from one of the ex-



(a) **Transferability** across various ResNet backbone **depths** originally trained using the ArcFace head.



(b) **Transferability** across various ResNet backbones with a fixed depth of 100 layers which were learned using different **heads**.

Figure 5: Transferability experiments measured in terms of cosine similarity and visualized using a heat map. Rows are divided into three groups: control masks (top four rows), mask trained using a single model (four next rows), mask trained using all four models (last row).

(a) (b) (c)

Figure 6: An illustration of our adversarial masks trained using the ResNet100 backbone and the following heads: (a) ArcFace (b) CosFace, and (c) MagFace.

amined datasets (presented earlier in this section) and study its effectiveness on the other datasets (i.e., the ground-truth embedding vectors are generated using other dataset's images). For simplicity's sake, we train all of the masks using

	CASIA	CelebA	MS-Celeb
CASIA	0.106	0.128	0.114
CelebA	0.107	0.095	0.103
MS-Celeb	0.118	0.116	0.106

Table 2: Cosine similarity score when training the mask using images from one dataset and testing on embedding vectors generated using images from other datasets.

the ResNet100 backbone originally trained with the ArcFace head. As shown in Table 2, the use of a specific dataset is insignificant, since our mask generalizes over all datasets.

Female vs. male adversarial mask. Another aspect we studied is the effect of a specific gender on the learned mask. We conducted two experiments in which the mask was trained using female or male facial images only. The results show that even when training the mask on a specific gender, the cosine similarity decreases to the same level as the trained mask on both genders (~ 0.1). In addition, masks learned by a single gender were able to transfer very well to the other gender. Generally, the contour of the learned masks (including the mask learned on both genders) is quite interesting. The masks resemblance to a human face "peels" another layer towards better explainability of these DNN-based models. More specifically, the resemblance of all learned masks to a male face might indicate there is an underlying bias hidden in these models.

4.1.2 Physical Attacks

Finally, to evaluate the effectiveness of our attack in the real world, we print our digital pattern on two surfaces: on regular paper cut in the shape of a face mask and on a white fabric mask, as shown in Figure 7. In addition, we create a testbed that operates an end-to-end face recognition system (as explained in Section 2), simulating a CCTV use case.

The system contains: (a) a *Dahua IPC-HDBW1431E* network camera which records a long corridor, (b) an MTCNN [44] detection model for face detection, preprocessing, and alignment, and (c) an attacked model - we perform a white-box attack in which the model used for training the adversarial mask is also the model under attack, a ResNet100 backbone originally trained using the ArcFace head. We use cosine similarity as the similarity metric.

To calculate the specific verification threshold (set at 0.38), we use a subset of 1,000 identities from the CASIA-WebFace dataset and perform the following procedure. Various face masks are applied (digitally) to each identity's original facial images. Then, we calculate the cosine similarity between the identity's embedding vector and each masked face image. Since we employ a semi-critical security use case (CCTV), we chose the threshold that led to a false acceptance rate (FAR) of 1%. Furthermore, to minimize false positive alarms, we used a persistence threshold of $N_{\text{recognized}} = 7$ frames and a sliding window of

Mask Type	Recognition Rate (RR)	Cosine Similarity	Persistence Detection
Clean	74.83%	0.52	100%
Blue	53.04%	0.401	100%
Random	54.76%	0.412	100%
Female Face	30.85%	0.308	96.67%
Male Face	28.36%	0.311	83.34%
Adv Fabric	5.72%	0.226	3.34%
Adv Paper	4.61%	0.181	0%

Table 3: Physical experiments' averaged results on all participants across different evaluated masks.

$N_{\text{sliding window}} = 10$ frames to designate a candidate identity as a valid one.

A group of 15 male and 15 female participants was assembled (after approval was granted by the university's ethics committee). Each participant was asked to walk along the corridor seven times, once for each mask evaluated (clean, blue, random, male face, and female face), similar to the digital experiments, and two more times with our adversarial masks printed on paper and fabric. The ground-truth embedding of each participant was calculated using two facial images, where a standard face mask was applied (digitally) to each image, for a total of four facial images.

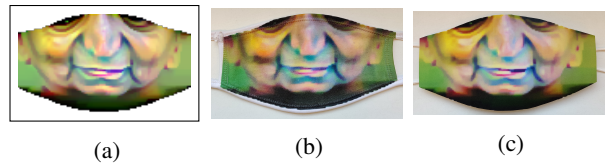
A demo of our experiments can be found here: https://youtu.be/_TXkDO5z11w.

The results of our experiments are shown in Table 3 where we can see that our adversarial masks (paper and fabric) performed significantly better than the other masks evaluated on every metrics, with a high correlation to the cosine similarity results obtained in the digital domain.

In terms of the RR, the performance of the FR model on the different masks can be divided into four groups (decreasing order): (a) the unmasked version (74.83%), (b) blue and random masks (53.04% and 54.76%, respectively), (c) male and female masks (30.85% and 28.36%, respectively), and (d) our fabric and paper adversarial masks (5.72% and 4.61%, respectively).

In a realistic case of CCTV use in which an attacker tries to evade the detection of the system, our adversarial fabric mask was able to conceal the identity of 29 out of 30 participants (which represents a persistence detection value of 3.34%), as opposed to the control masks which were able to decrease it to only 83.34% at most.

Another aspect we examined in our physical evaluation is the ability to print the adversarial pattern on a real surface. Figure 7b and 7c present the digital adversarial pattern (7a) printed on the different surfaces. Due to the limited ability of a printer to accurately output the original colors onto the fabric, we can see that there is a slight difference in the performance of the paper and fabric masks. Nonetheless, both of our adversarial masks outperformed the other masks evaluated.



(a) (b) (c)

Figure 7: An illustration of: (a) the digital adversarial mask trained on the ResNet100 backbone with the ArcFace head; (b) the digital pattern printed on fabric mask; and (c) the digital pattern printed on paper.

5. Countermeasures

We propose two ways in which our digital masking method can be used to deal with such adversarial attacks: (a) adversarial training – adversarial (universal and tailor-made) masked face images could be provided to the trained model to improve its robustness; and (b) mask substitution – during the inference phase, every masked face image could be preprocessed so that the mask applied is replaced digitally with a standard one (e.g., blue mask 3b), where the models had satisfactory performance, as shown in Section 4, eliminating the potential threat of an adversarial face mask. An implementation of this method on facial images of 100 identities ($\sim 10k$ images) from the CASIA-WebFace dataset increased the RR from 0.4% (the adversarial mask is applied to the facial images) to 65.5% (the blue mask is applied to the adversarial images). We also implemented the proposed method in a physical experiment in which the blue disposable mask was digitally placed on facial images extracted from the videos frames (videos of participants wearing the adversarial mask), improving the RR from 5.72% to 57.3%.

6. Conclusion

In this paper, we showed how an off-the-shelf face mask covered with a carefully crafted adversarial pattern is able to prevent an FR system from identifying a potential attacker wearing the mask. Whereas other attack methods used different accessories that are more conspicuous and do not blend naturally in the environment, our mask will not raise any suspicion due to the widespread use of face masks during the COVID-19 pandemic. We demonstrated the effectiveness of our mask in the digital domain, both under white-box and black-box conditions. In the white-box setting our mask outperformed all evaluated masks and transferred well to other models (black-box setting). In the physical domain, we showed how our mask is able to prevent the detection of multiple participants in a CCTV use case system. Moreover, we proposed possible countermeasures to deal with such attacks during both the training and inference phases. To sum up, in this research, we highlight the potential risk FR models face from an adversary simply wearing a face mask in the COVID-19 era.

References

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Fu Ying. Partial fc: Training 10 million identities on a single machine. In *Arxiv 2010.05222*, 2020. 4

[2] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 4

[3] Cunjian Chen. Pytorch face landmark: A fast and accurate facial landmark detector, 2021. Open-source software available at https://github.com/cunjian/pytorch_face_landmark. 5

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 5

[5] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 2

[6] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 2, 3, 4

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 1, 2

[9] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 3

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[11] Nitzan Guetta, Asaf Shabtai, Inderjeet Singh, Satoru Momiyama, and Yuval Elovici. Dodging attack using carefully crafted natural makeup. *arXiv preprint arXiv:2109.06467*, 2021. 1, 2

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 720–729, 2020. 2

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021. 1, 2

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 1, 2

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[19] Juncheng Li, Frank R Schmidt, and J Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. *arXiv preprint arXiv:1904.00759*, 2019. 2

[20] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018. 2

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 4

[22] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 3, 4

[23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 1, 2

[25] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 2

[26] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2

[28] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 1, 2, 4

[29] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019. 2

[30] Meng Shen, Zelin Liao, Liehuang Zhu, Ke Xu, and Xiaojiang Du. Vla: A practical visible light-based attack on face recognition systems in physical world. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–19, 2019. 1, 2

[31] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018. 1, 2

[32] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 1, 2

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[34] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2

[35] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 3, 4

[36] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. *arXiv preprint arXiv:2101.04407*, 2021. 3, 5

[37] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. *arXiv preprint arXiv:1910.14667*, 2019. 2

[38] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 4

[39] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. *arXiv*, pages arXiv–1910, 2019. 2

[40] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, pages 665–681. Springer, 2020. 4

[41] Xiao Yang, Fangyun Wei, Hongyang Zhang, and Jun Zhu. Design and interpretation of universal adversarial patches in face detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 174–191. Springer, 2020. 1, 2

[42] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 4

[43] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021. 2

[44] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 7

[45] Alon Zolfi, Moshe Kravchik, Yuval Elovici, and Asaf Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15232–15241, 2021. 1, 2