Adversarial Mask: Real-World Universal Adversarial Attack on Face Recognition Models

Alon Zolfi[1][0000−0003−0270−1743]✉, Shai Avidan[2],
Yuval Elovici[1][[0000−0002−9641−128X], and Asaf Shabtai[1][0000−0003−0630−4059]

[1] Ben-Gurion University of the Negev, Israel
zolfi@post.bgu.ac.il, {elovici,shabtaia}@bgu.ac.il
[2] Tel Aviv University, Israel
avidan@tauex.tau.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, and thus may raise suspicion in real-world environments (e.g., airports). In this paper, we propose *Adversarial Mask*, a physical universal adversarial 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's effectiveness in real-world experiments (CCTV use case) by printing the adversarial pattern on a fabric face mask. In these experiments, the FR system was only able to identify 3.34% of the participants wearing the mask (compared to a minimum of 83.34% with other evaluated masks). A demo of our experiments can be found at: https://youtu.be/_TXkDO5z11w.

Keywords: Adversarial Attack · Face Recognition · Face Mask

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 facial recognition (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.

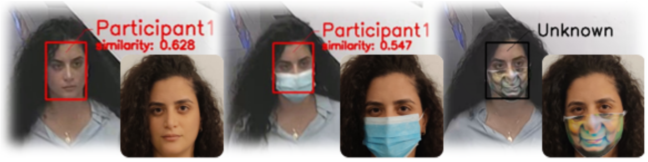


Fig. 1: Illustrating the effect of an adversarial pattern printed on a fabric mask (right), which results in the failure of the FR system to detect the person wearing it, compared to the FR system's ability to detect the same individual without a mask, as well as with a standard disposable mask.

Adversarial attacks in the computer vision domain have gained a lot of interest in recent years, and various ways of fooling image classifiers [9, 22] and object detectors [21, 23, 32] 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 [28]. In the physical domain, some of the proposed methods involved wearing adversarial eyeglasses [18], projecting lights on human faces [20], wearing a hat containing an adversarial sticker [14], and using adversarial makeup [10]. 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 aims to cause *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 evaded the detection by the FR system.

The contributions of our research can be summarized as follows:

- We are the first to present a **physical universal** adversarial attack that fools FR models, i.e., we craft a single perturbation that causes the FR model to falsely classify all potential attackers as unknown identities, even under diverse conditions (angles, scales, etc.) in a real-world environment (fully-automated CCTV scenario).
- In the digital domain, we study the transferability of our attack across different model architectures and datasets.
- We present a fully differentiable 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 craft an inconspicuous pattern that "continues" the contour of the face, allowing a potential attacker to easily blend in with a crowd without raising an alarm, given the variety and widespread use of face masks during the COVID-19 pandemic.
- We propose various countermeasures that can be used during the FR model training and inference phases.

2 Background & Related Work

2.1 Adversarial Attacks

Digital Attacks. Initially, attacks in the digital domain aimed at fooling classification models were introduced [9, 22]. While those earlier attacks are based on methods that generate a perturbation for a single image, Moosavi-Dezfooli *et al.* [17] proposed universal adversarial perturbations (UAPs), which enable 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. Yang *et al.* [28] designed a digital patch which is placed on a person's forehead to deceive face detectors. Recent studies targeting FR models suggested various techniques. Deb *et al.* [6] proposed automated adversarial face synthesis, using a generative adversarial network (GAN) to create minimal perturbations. Agarwal *et al.* [1] and Amada *et al.* [2] proposed UAPs that can deceive FR models for multiple identities simultaneously. However, these attacks only call attention to the potential threat inherent to such models but cannot be transferred to the physical world.

Physical Attacks. Physical attacks differ from digital attacks in the way real-world constraints are considered throughout the process of generating the perturbation. Consequently, these constraints allow the perturbations to transfer more easily to the physical world. In recent years, physical attacks on object detectors have gained attention. Chen *et al.* [5] printed stop signs containing adversarial patterns that evaded detection by the object detector, and Sitawarin *et al.* [21] 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.* [23] suggested attaching a small adversarial cardboard plate to a person's body to evade detection. 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 into the environment more naturally [26, 27]. A slightly different approach, in which the perturbation affects the sensor's perception of the object by applying a translucent patch on the camera's lens, was also introduced [32].

Numerous studies have demonstrated different ways of fooling FR systems. For example, Shen *et al.* [20] 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 [10, 30].

Accessories were also shown to be effective; for example, Sharif *et al.* [18] 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 [19]. Recently, Komkov *et al.* [14] printed an adversarial paper sticker and placed it on a hat to fool the state-of-the-art *ArcFace* [7] FR model. However, 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 craft tailor-made perturbations (target a single image or person), our universal attack can be applied more widely without the need for an expert to train a tailor-made one. Furthermore, we demonstrate the effectiveness of our method in a real-world use case involving a CCTV system, an aspect not addressed by previous studies.

2.2 Face Recognition

Models. FR models can be categorized by two main attributes, the model's backbone and the novel loss function, both of which are involved in the training phase. The main architecture used as the backbone in these models is the ResNet [12] architecture, which varies in terms of the number of layers it contains, also referred to as the backbone *depth*. On top of the backbone, an additional layer (or more) is added, usually containing a novel loss function that is used to train the backbone weights [7, 16, 24]. Later, when the FR model is used for inference, only the backbone layers are used to generate the embedding vector.

Systems. The end-to-end procedure of a fully automated FR system consists of several main steps: (a) Record - a camera records the environment and then produces a series of frames (a video stream); (b) Detect - each frame is analyzed by a face detector to extract cropped faces; (c) Align - the cropped faces are aligned according to the FR model's alignment method; (d) 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; (e) Verify - the embedding vector is compared to a list of precalculated embedding vectors (also referred to as ground-truth embedding vectors) using a similarity measure (e.g., cosine similarity). The identity with the highest similarity score is marked as a potential candidate and eventually confirmed if its similarity score surpasses a predefined verification threshold (which depends on the system's use case).

3 Method

The objective of our research is to generate an adversarial pattern that can be printed on a face mask and cause FR systems to classify a registered identity as an unknown 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 loss functions).

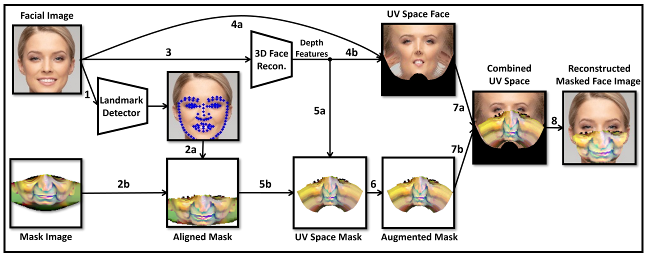


Fig. 2: Overview of our mask projection method pipeline.

3.1 Mask Projection

In order to digitally train our adversarial mask, we first need to simulate the mask 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.* [8] 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 a 3D face and provides dense correspondence to the semantic meaning of each point in the UV space, allowing us to achieve near-real approximation of the mask on the face, which is essential to the creation of a successful adversarial mask in the real world.

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 [25]) takes a mask M_{adv} and a facial image x_{face}, and applies the mask on the face, resulting in a masked face image $\mathcal{R}_\theta(M_{\text{adv}}, x_{\text{face}})$. As shown in Figure 2, the pipeline of the mask's projection on the facial image is as follows:

1. Detect the landmark points of the face - given a landmark detector, we extract the landmark points of the face.
2. Map the mask pixels to the facial image - the landmark points of the face extracted in the previous step of the pipeline are used to map the mask pixels to the corresponding location on the facial images.
3. Extract depth features of the face - the facial image is passed to the 3D face reconstruction model to obtain depth features.
4. Transfer 2D facial image to the UV space - the depth features are used to remap the facial image to the UV space.
5. Transfer 2D mask image to the UV space - the depth features are used to remap the mask image to the UV space.

6. Augment mask - to improve the robustness of our adversarial mask, random geometric transformations and color-based augmentations (parameterized by θ) are applied: (i) geometric transformations - random translation and rotation are added to simulate possible distortions in the mask's placement on the face in the real world, and (ii) color-based augmentations - random contrast, brightness, and noise are added to simulate changes in the appearance of the mask that might result from various factors (e.g., lighting, noise or blurring caused when the camera captures the image).
7. Combine and reconstruct - the UV representations of the facial image and the mask are combined, and the combined image is reconstructed back to the regular 2D space, resulting in a masked face image.

Usually, adversarial attacks that employ textile-like objects (e.g., wearable t-shirt [26,27]) use thin plate splines (TPSs) [4] 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

To optimize our mask's pixels, we propose an iterative optimization process. In each iteration, we select a random batch of facial images of multiple identities and digitally project the mask on each facial image. We then feed the masked face images to the FR model and obtain the embedding representations. Since our goal is to cause an attacker to be unknown to FR models, we aim to create a patch M_{adv} that will decrease the similarity between the output embedding and the ground-truth embedding e_{gt} (precalculated) for each identity.

More formally, an FR model $f : \mathcal{X}^{w \times h \times 3} \rightarrow \mathbb{R}^N$ receives a facial image $x \in \mathcal{X}$ (in our case, a masked face image $\mathcal{R}_\theta(M_{adv}, x)$) as input and outputs the embedding representation $f(\mathcal{R}_\theta(M_{adv}, x))$. Therefore, we minimize the cosine similarity between the embedding vectors and use the following loss function:

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

Since our method is not system-dependent (i.e., does not use a 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, denoted as J. We replace 1 with the following:

$$\ell_{sim}(M_{adv}) = \mathbb{E}_{\theta,x} \frac{1}{|J|} \sum\nolimits_j \cos(f^{(j)}(\mathcal{R}_\theta(M_{adv}, x)), e_{gt}^{(j)}), \tag{2}$$

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)* [18] factor to ensure that the optimizer favors smooth color transitions between neighboring pixels and is calculated on the mask pixels as follows:

$$\ell_{TV} = \sum\nolimits_{i,k} \sqrt{(p_{i,k} - p_{i+1,k})^2 + (p_{i,k} - p_{i,k+1})^2} \tag{3}$$

When neighboring pixels are not similar, the penalty of this component is greater.

To be more precise, since the output of ℓ_{sim} is in the range of $[-1, 1]$ and the output of ℓ_{TV} is in the range of $[0, 1]$, we transform ℓ_{sim} so it is in the same range ($[0, 1]$); thus, we replace 2 with the following:

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

Finally, the optimization problem we solve is as follows:

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

where λ is set at a low 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. We use three different types of loss functions that were originally used to train the models, which are considered state-of-the-art: ArcFace [7], CosFace [24], and MagFace [16]. Specifically, we use pretrained models which were trained using the ArcFace and CosFace loss functions [3], with four different ResNet depths (18, 34, 50, and 100) each, and a pretrained ResNet100 backbone originally trained with the MagFace [16] loss function, 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 in a 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 commonly used datasets in the face recognition domain: CASIA-WebFace [29], CelebA [15], and MS-Celeb [11].

For the training phase, we randomly choose 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 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 decrease the similarity score - specifically the cosine similarity (an approach originally presented in [14]). The cosine similarity calculation is a step required prior to making a binary decision based on a predefined threshold. This evaluation approach does not require a system-dependent predefined threshold and demonstrates our attack's effectiveness. 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:

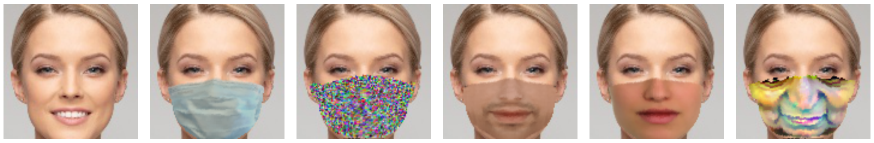
- *Recognition rate* (RR) $= |F_{rec}|/|F_{det}|$, where $|F_{rec}|$ denotes the total number of frames in which the identity was correctly recognized (the cosine similarity between the ground-truth embedding and the output embedding surpasses the verification threshold), and $|F_{det}|$ denotes the 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 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, and we set the initial color of the mask to white. The pixels are updated using the Adam optimizer [13], where the initial learning rate is set at 10^{-2}. The weight factor of the TV component in the loss function λ is manually set at 0.1. The source code is available online.[3]

Types of face masks evaluated. Since we are the first to present a physical universal perturbation, we compare the effectiveness of our mask with several control masks: (a) Clean - the original facial image without a mask, (b) Adv - our optimized adversarial mask, (c) Random - a mask with randomly colored pixels, and (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.

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 (ResNet100@ArcFace) performance on a number of simple face masks. For this evaluation, we use 100 identities from the CASIA-WebFace dataset, where five images of each identity are used to calculate the ground-truth embedding, and the remaining images are applied with different types of masks.

[3] https://github.com/AlonZolfi/AdversarialMask



(a) Clean (b) Blue (c) Random (d) Male (e) Female (f) Adv

Fig. 3: Examples of facial images w/o mask (a), and when various masks are digitally applied to them (b)-(f).

Table 1: Cosine similarity comparison between two ground-truth embedding generation methods on the Resnet100@ArcFace. Bold indicates better performance.

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

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

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 use two approaches for 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 of the first approach - in addition to the original images, we create a masked face version for each image (the specific mask is randomly chosen from blue, black, and white masks) and average the embedding vectors of the two versions of the images. We then calculate the cosine similarity between the masked face images' embedding vectors and the two versions of ground-truth embedding vectors generation.

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 vectors used for the training procedure are generated using first approach). 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 face mask.

4.1 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.

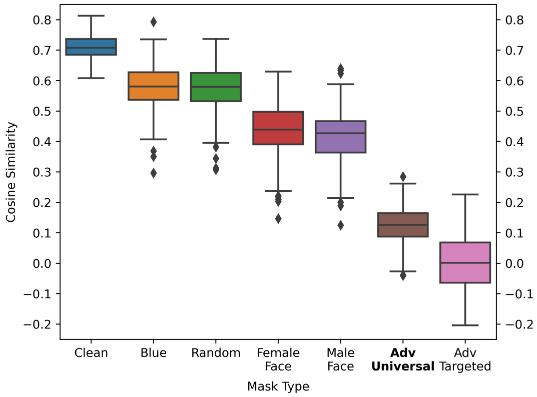
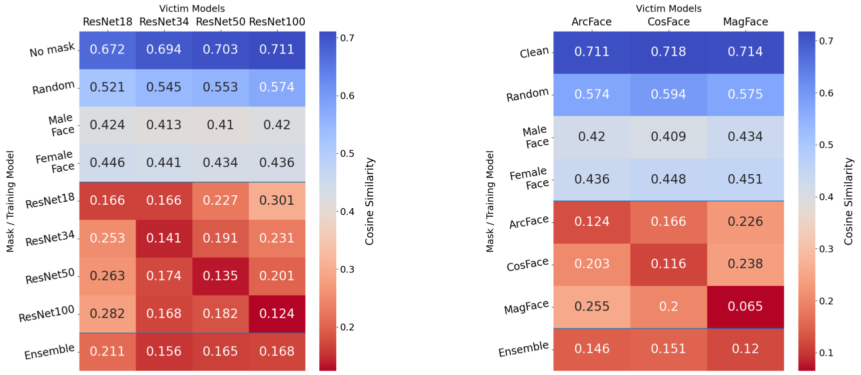


Fig. 4: Distribution of the cosine similarity score across different masks. ’Adv Universal’ represents our optimized universal mask, and ’Adv Targeted’ represents a tailor-made mask for each identity.

Effectiveness of the adversarial mask in a 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@ArcFace. As shown in Figure 4, our adversarial mask has a significant impact compared to 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 perform 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 represent an attack that is more difficult to detect, since the adversarial pattern varies among different identities. In addition, while the female and male face control masks are also able to decrease the cosine similarity to a lower level (~ 0.45), our mask outperforms them both for almost all tested identities.

Transferability across backbone depth. We also 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 the ResNet@ArcFace. The results are presented in Figure 5a. In the figure, 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 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 a model with 100 layers performs better on the models with 34 and 50 layers (decreasing the cosine similarity to 0.182 and 0.168, respectively) than on the 18-layer model (0.282). In addition, we see that the mask trained on an ensemble of



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

(b) **Transferability** across various ResNet100 backbones originally trained with different **loss functions**.

Fig. 5: Transferability experiments measured in terms of cosine similarity. Rows are divided into three groups: control masks, masks trained using a single model, mask trained using all of the models.

all models does not outperform a mask trained on a single model in a white-box setting, however the ensemble's effectiveness is seen over all models combined.

Transferability across different loss functions. We further demonstrate the adversarial mask's transferability across different model loss functions. We use the ResNet100 backbone in which the weights were trained using one of the following loss functions: ArcFace, CosFace, and MagFace. In Figure 5b, we observe that our method is loss-agnostic, as the decrease in the cosine similarity is seen on for all tested models. However, a mask that was trained using the MagFace model does not generalize as well as the masks trained with 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 models, respectively. It is interesting to examine the mask trained by each model (presented in Figure 6). Whereas there is a resemblance in the contour of the optimized masks, the mask trained using the ResNet100@MagFace backbone (Figure 6c) learns completely different colors than the other two, in some way providing a possible explanation for its decreased ability to generalize to the ArcFace and CosFace models.

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 examined datasets (presented earlier in this section) and study its effectiveness on the other datasets (i.e., the ground-truth embedding vectors are generated using another dataset's images). We train all of the masks using the ResNet100@ArcFace. The results show that the impact of using a specific dataset is insignificant, since our mask generalizes over all datasets. For example, when

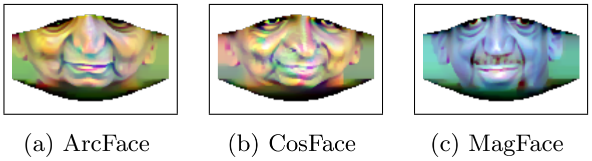

(a) ArcFace (b) CosFace (c) MagFace

Fig. 6: Illustrations of our adversarial masks trained on different ResNet100 backbones, which vary in terms of the original loss function they were trained on.

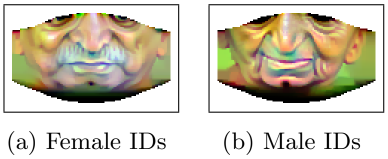

(a) Female IDs (b) Male IDs

Fig. 7: The adversarial masks trained on the ResNet100@ArcFace using single gender identities.

training the mask on the CASIA-WebFace dataset and testing it on the CelebA and MS-Celeb datasets, we respectively obtained an average cosine similarity of 0.128 and 0.114, similar to the white-box setting results (mask trained and tested on images from the CASIA-WebFace dataset, Figure 4).

Effect of gender. Another aspect we studied is the effect of a specific gender on the trained mask. The experiments include optimization of the adversarial mask using only female or male identities, and the final masks are presented in Figure 7a and Figure 7b, respectively. The results show that even when training the mask on facial images of a single gender, the cosine similarity decreases to the same level as the mask trained on both genders (~ 0.1). In addition, masks trained by a single gender were able to transfer very well to the other gender (male $\rightarrow$ female $= 0.097$, and female $\rightarrow$ male $= 0.145$).

Generally, the contour of the trained masks (including the mask trained on both genders, Figure 6a) is quite interesting. Despite the fact that only facial images of female identities were used to train the mask (Figure 7a), the optimized mask has an high resemblance to a male face. More generally, the resemblance of all the trained masks to a male face might indicate there is an underlying bias hidden in these models.

4.2 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 9. In addition, we create a testbed that operates an end-to-end fully automated FR system (explained in Section 2), simulating a CCTV use case.

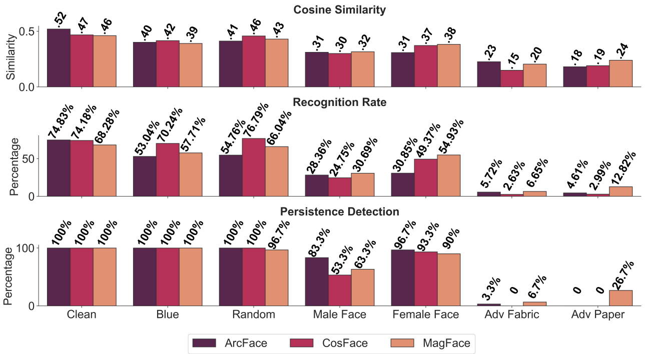


Fig. 8: Physical experiments' averaged results on all participants across different evaluated masks and different victim models.

Setup. The system contains: (a) a *Dahua IPC-HDBW1431E* network camera which records a long corridor, (b) an MTCNN [31] 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@ArcFace. In addition, we perform an "offline" analysis in a black-box setting, in which the facial images are cropped from the original frames and compared to ground-truth embedding vectors generated using other models.

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

We recruited a group of 15 male and 15 female participants (after approval was granted by the university's ethics committee). Each participant was asked to walk along the corridor seven times, once with 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.

Results. The results of our experiments are shown in Figure 8 where we can see that our adversarial masks (paper and fabric) performed significantly better than the other masks evaluated on every metric, with a high correlation to the cosine similarity results obtained in the digital domain.

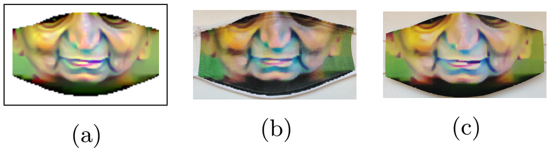

(a) (b) (c)

Fig. 9: An illustration of: (a) the digital adversarial mask trained on the ResNet100@ArcFace; (b) the digital pattern printed on fabric mask; and (c) the digital pattern printed on paper.

In terms of the RR, the performance of the FR model for the different masks can be divided into four groups (listed in 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 conceal 5 out of 30 participants at most (persistence detection value of 83.34%).

We also examine the effectiveness of our masks on models they were not trained on. The results presented in Figure 8 show that our masks have similar adversarial effect on FR models in a black-box setting as in a white-box setting.

Another aspect we examined in our physical evaluation is the ability to print the adversarial pattern on a real surface. Figures 9b and 9c present the digital adversarial pattern (9a) 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 masks. Nonetheless, both of our adversarial masks outperformed the other masks evaluated.

5 Countermeasures

We propose two ways in which our digital masking method can be used to defend against adversarial masks: (a) adversarial training – adversarial (universal and tailor-made) masked face images could be provided to the model during training to improve its robustness; and (b) mask substitution – during the inference phase, every masked face image could be preprocessed so that the worn mask 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 the mask substitution 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). In a physical experiment, in which the blue mask was digitally placed on facial images extracted from the videos frames (videos of participants wearing the adversarial mask), the RR increased from 5.72% to 57.3%.

6 Conclusion

In this paper, we presented a physical universal attack in the form of a face mask against FR systems. 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 settings. 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. To sum up, in this research, we highlight the potential risk FR models face from an adversary simply wearing a carefully crafted adversarial face mask in the COVID-19 era.

References

1. Agarwal, A., Singh, R., Vatsa, M., Ratha, N.: Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–7. IEEE (2018)
2. Amada, T., Liew, S.P., Kakizaki, K., Araki, T.: Universal adversarial spoofing attacks against face recognition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–7. IEEE (2021)
3. An, X., Zhu, X., Xiao, Y., Wu, L., Zhang, M., Gao, Y., Qin, B., Zhang, D., Ying, F.: Partial fc: Training 10 million identities on a single machine. In: Arxiv 2010.05222 (2020)
4. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence **11**(6), 567–585 (1989)
5. Chen, S.T., Cornelius, C., Martin, J., Chau, D.H.P.: Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 52–68. Springer (2018)
6. Deb, D., Zhang, J., Jain, A.K.: Advfaces: Adversarial face synthesis. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2020)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
8. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 534–551 (2018)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

10. Guetta, N., Shabtai, A., Singh, I., Momiyama, S., Elovici, Y.: Dodging attack using carefully crafted natural makeup. arXiv preprint arXiv:2109.06467 (2021)
11. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Komkov, S., Petiushko, A.: Advhat: Real-world adversarial attack on arcface face id system. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 819–826. IEEE (2021)
15. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset. Retrieved August **15**(2018), 11 (2018)
16. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14225–14234 (2021)
17. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
18. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: 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. pp. 1528–1540 (2016)
19. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security (TOPS) **22**(3), 1–30 (2019)
20. Shen, M., Liao, Z., Zhu, L., Xu, K., Du, X.: 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)
21. Sitawarin, C., Bhagoji, A.N., Mosenia, A., Chiang, M., Mittal, P.: Darts: Deceiving autonomous cars with toxic signs. arXiv preprint arXiv:1802.06430 (2018)
22. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
23. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
24. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
25. Wang, J., Liu, Y., Hu, Y., Shi, H., Mei, T.: Facex-zoo: A pytorch toolbox for face recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3779–3782 (2021)
26. Wu, Z., Lim, S.N., Davis, L.S., Goldstein, T.: Making an invisibility cloak: Real world adversarial attacks on object detectors. In: European Conference on Computer Vision. pp. 1–17. Springer (2020)
27. Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., Chen, P.Y., Wang, Y., Lin, X.: Adversarial t-shirt! evading person detectors in a physical world. In: European Conference on Computer Vision. pp. 665–681. Springer (2020)

28. Yang, X., Wei, F., Zhang, H., Zhu, J.: 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. pp. 174–191. Springer (2020)
29. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
30. Yin, B., Wang, W., Yao, T., Guo, J., Kong, Z., Ding, S., Li, J., Liu, C.: Adv-makeup: A new imperceptible and transferable attack on face recognition. arXiv preprint arXiv:2105.03162 (2021)
31. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
32. Zolfi, A., Kravchik, M., Elovici, Y., Shabtai, A.: The translucent patch: A physical and universal attack on object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15232–15241 (2021)