Defending Black Box Facial Recognition Classifiers Against Adversarial Attacks

Rajkumar Theagarajan and Bir Bhanu
Center for Research in Intelligent Systems, University of California, Riverside, CA 92521
rthea001@ucr.edu, bhanu@cris.ucr.edu

Abstract

Defending adversarial attacks is a critical step towards reliable deployment of deep learning empowered solutions for biometrics verification. Current approaches for defending Black box models use the classification accuracy of the Black box as a performance metric for validating their defense. However, classification accuracy by itself is not a reliable metric to determine if the resulting image is "adversarial-free". This is a serious problem for online biometrics verification applications where the ground-truth of the incoming image is not known and hence we cannot compute the accuracy of the classifier or know if the image is "adversarial-free" or not. This paper proposes a novel framework for defending Black box systems from adversarial attacks using an ensemble of iterative adversarial image purifiers whose performance is continuously validated in a loop using Bayesian uncertainties. The proposed approach is (i) model agnostic, (ii) can convert single step black box defenses into an iterative defense and (iii) has the ability to reject adversarial examples. This paper uses facial recognition as a test case for validating the defense and experimental results on the MS-Celeb dataset show that the proposed approach can consistently detect adversarial examples and purify/reject them against a variety of adversarial attacks with different ranges of perturbations.

1. Introduction

Deep learning has achieved impressive performance [5, 29, 33, 34], significantly improving the development for a variety of biometrics applications [4, 6, 46, 48, 49, 55, 59, 67, 72]. However, it has been shown that deep learning classifiers are vulnerable to adversarial attacks which are intentionally designed to cause misclassifications [30, 37, 60]. These attacks are carefully crafted perturbations, added to an image that are visually imperceptible to the human eye, and they can cause deep learning models to misclassify the image with high confidence. In the domain of adversarial attacks, there are two types of threat models: 1) White box, and 2) Black box attacks. In the white box setting [10], the attacker has full knowledge about the classification model's parameters and architecture, whereas in the Black box setting [50] the attacker does not have this knowledge. In this paper we focus on the Black box based adversarial attacks.

Current defenses against adversarial attacks can be classified into four approaches: 1) modifying the training data, 2) modifying the model, 3) using auxiliary tools, and 4) detecting and rejecting adversarial examples. Modifying the training data involves augmenting the training dataset with adversarial examples and re-training the classifier [22, 30, 64, 73] or performing *N* number of pre-selected image transformations in a random order [13, 17, 26, 52]. Modifying the model involves pruning the architecture of the classifier [38, 51, 69] or adding pre/post-processing layers to it [9, 12, 71]. These approaches are not compatible to be used for defending Black box models. Using auxiliary tools involves having an independent module that is able to process the input before it is passed to the classifier [43, 57, 62**?**]. Detecting and rejecting adversarial examples involve using domain adaptation techniques and carrying out statistical analysis [20, 25, 40] to detect adversarial examples. The approaches in the latter two categories are the most suitable for defending Black box classifiers. A drawback of these approaches is that, they do not quantify how much adversarial component is left in the resulting purified image and they perform well only under weakly bounded adversarial perturbations. This is very important, especially for online biometrics verification systems, because an adversary can always use an attack with a comparatively higher perturbation that is just enough to fool the system. Even though it is trivial for an attacker to increase the adversarial perturbation as they are relatively more discernible to the human eye, this is not the case for safety-critical applications that do not have a human in the loop. Additionally, the adversary can also use physical adversarial attacks as shown in [36, 63, 70] which are not very obvious to the human eye. Thus, in biometrics verification approaches where there is no human intervention, it is very important to quantify/explain the adversarial component in the resulting image of any adversarial defense.

To overcome this problem, *this is the first paper that pro-*

Name	Domain	Comments
Adversarial attacks in face biometrics		
Goswami *et al.* [23]	Black box	Used image and face level distortions to attack face recognition systems
Bosse *et al.* [8]	White box	Used adversarial transformers and attacked Faster RCNN to obfuscate the person
Dong *et al.* [16]	Black box	Used evolutionary algorithms to create adversarial examples
Lu *et al.* [41]	White box	Used FGSM attack to create a single noise for all frames in a video
Milton [44]	Black box	Used a momentum based FGSM attack and knowledge distillation to attack CNNs
Zhou *et al.* [75]	White box	Used infrared LEDs attached to a head wear to create physical adversarial examples
Deb *et al.* [14]	White box	Used GAN [21] to generate adversarial perturbations in subtle facial regions
Adversarial defenses in face biometrics		
Wadhwa *et al.* [65]	Black box	Used GAN [21] to detect adversarial authentications using differences between the classification and matching thresholds
Agarwal *et al.* [2]	Black box	Used Haralick texture features from discrete wavelet transformed frames for classification using PCA + SVM
Goswami *et al.* [24]	White box	Used the representation of the hidden layers to detect adversarial attacks and performed selective dropout on the affected filters.
Agarwal *et al.* [1]	Black box	Used transformations such as gamma correction, log transform, and brightness control to fool face presentation attack detection algorithms.
Tao *et al.* [61]	White box	Used activation of hidden layers to determine if the CNN is looking at facial regions
Agarwal *et al.* [3]	White box	Used PCA + SVM to detect the presence of universal attack

Table 1: State-of-the-art of adversarial attacks and defenses in the domain of face biometrics

poses to use an ensemble of iterative adversarial image purifiers whose performance is validated by quantifying the remaining amount of adversarial component after each iteration of purification using Bayesian uncertainties. By doing so, the proposed approach is able to quantify after each iteration if the resulting image is adversarial or not without the need for any ground-truth/human observer. Experimental results show that using an ensemble of adversarial defenses performs much better than using a single stand-alone defense. It should be noted that the scope of this paper is not to design a robust facial recognition classifier, but instead, to design a robust defense to protect existing facial recognition classifiers from adversarial attacks.

2. Adversarial Defenses and Attacks in Face Biometrics

Table 1 shows a brief summary of the adversarial defense and attacks used in the field of face biometrics. In summary the contributions of this paper are:

- Defending Black box classifiers using an ensemble of independent iterative adversarial image purifiers.
- Validating the performance of the proposed defense continuously in a loop using Bayesian uncertainties and classification accuracy.
- The proposed approach is model agnostic and can convert any single step adversarial image purifier into an iterative adversarial image purifier.

3. Technical Approach

In this section we explain the individual modules of our approach shown in Fig. 1. The input image (X) first passes through the Bayesian CNN and if the image is adversarial, it is purified by the ensemble of defenses and the resulting purified image (X'_1) is passed as input back to the Bayesian CNN. If X'_1 is not adversarial it is passed as input to the Black box classifier. If X'_1 is still adversarial it is passed back to the ensemble of defenses and this continues for M iterations. After M iterations of purification, if X'_M is still adversarial then the image is rejected.

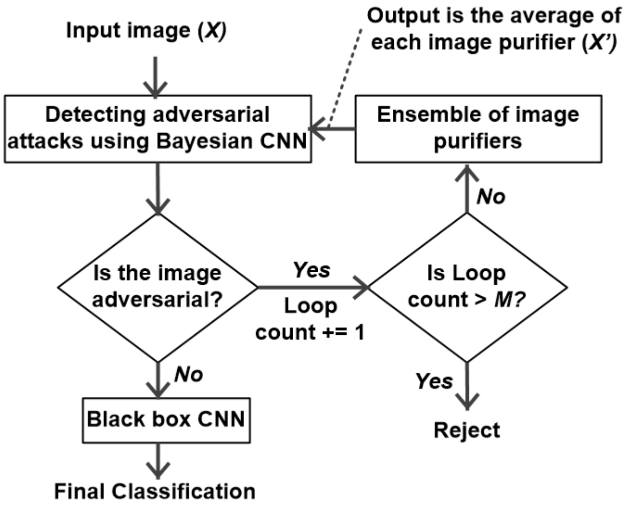

Input image (X)
Output is the average of each image purifier (X')
Detecting adversarial attacks using Bayesian CNN
Ensemble of image purifiers
No
Is the image adversarial?
Yes
Loop count += 1
Is Loop count > M?
No
Yes
Black box CNN
Reject
Final Classification

Figure 1: Overall framework of our approach.

· ***Assumptions of our Defense***: 1) The output of the Black box model is the top predicted class without any probabilities. 2) The architecture, parameters and *entire* training dataset of the Black box are not known to both the adversary and the defense algorithm. 3) The outputs of the Bayesian framework and ensemble of adversarial image purifiers are not shared with the adversary (i.e., the adversary can only see the output of the Black box).

· ***Target Applications for our Defense***: Our defense is suited for Black box applications that contain sensitive personnel information such as human biometrics for remote monitoring. These are critical applications that require security against adversarial attacks and preserve user privacy and do not want to give up sensitive information.

3.1. Threat Models

In this sub-section we define the adversarial attacks used for evaluating our defense. For a given test image-label pair (x, y), adversarial attacks find a perturbation δ with $||\delta||_\infty \leq \epsilon$ such that a deep learning classifier $f(\cdot)$ results in $f(x+\delta) \neq y$. ϵ is a hyper-parameter that sets the perturbation limit for each pixel in x on the color scale.

· **Fast Gradient Sign Method (FGSM)** [22]: This attack uses the sign of the gradients at every pixel to determine the direction of perturbation.

$$x^{adv} = x + \epsilon \cdot sign(\bigtriangledown_x L(x, y)) \qquad (1)$$

· **Basic Iterative Method (BIM)** [36]: This attack extends the FGSM attack [22] by iterating it multiple times with a small step size.

$$x^{adv}_{n+1} = Clip_\epsilon(x_n + \alpha \cdot sign(\bigtriangledown_x L(x_n, y))) \qquad (2)$$

· **Projected Gradient Descent** [42]: This attack computes the gradient in the direction of the highest loss and projects it back to the l_p norm around the sample.

$$x^{adv}_{n+1} = \prod^{\epsilon}(x_n + \alpha \cdot sign(\bigtriangledown_x L(x_n, y))) \qquad (3)$$

In eq. (1) - (3), $\bigtriangledown_x L(x, y)$ is the loss function used to train the CNN, α is the iterative step size, Clip $(\cdot)$ and $\prod(\cdot)$ are the clipping and projection functions, respectively.

3.2. Create a Substitute Model Using Knowledge Distillation

Since we do not know the architecture/parameters of the Black box CNN, we cannot directly apply our defense algorithm. To solve this we create a *Substitute* for the Black box model using *Knowledge Distillation*. *Knowledge Distillation* is the process of transferring functionality from a *Teacher* model to a *Substitute* model [11, 19, 47]. We assume the Black box CNN is deterministic (the weights do not change after training) and ignore approaches that incrementally update the weights of the CNN in an online manner (i.e., incremental [74] and reinforcement learning [53]).

Knowledge Distillation on the MS-Celeb [27] dataset: The MS-Celeb dataset consists of approximately 9.5M images for 99,892 celebrities. It has been shown that this dataset is extremely noisy with many incorrect annotations [15, 39, 66, 68]. This problem is pervasive since it is very time consuming to annotate very large scale datasets. In order to reduce the noise due to incorrect annotation, we followed the approach of Jin *et al*. [32]. The authors proposed a graph-based cleaning method that mainly employs the community detection algorithm [18] and deep CNN models to delete mislabeled images. Based on this method, the authors provided a list of correctly annotated images and showed that approximately 97.3% of images in the dataset are correctly labeled. This results in a total of approximately 6.5M images for 94,682 celebrities.

In order to train a Black box facial recognition classifier, we manually selected 100 celebrities that had at least 100 images after discarding images that had extremely skewed poses and celebrities wearing sunglasses. We denote this dataset as $Q_{1:100}$ (Q_i is the identity of the celebrity) and it is used for training the Black box classifier only. In order to train a *Substitute* model we need to first create a pseudo-labeled dataset. To do this, we probed the Black box CNN with images of celebrities that do not belong in $Q_{1:100}$ and labeled the images with the predicted class. We denote this dataset as $Q_{101:\infty}$ and it is used to train the *Substitute* model. It should be noted that the dataset $Q_{1:100}$ and $Q_{101:\infty}$ contain images of different celebrities and their data distributions do not overlap (ignoring the noise due to incorrect annotations). Although, we do not know the identities of the celebrities in $Q_{1:100}$, we assume that we have knowledge about the total number of celebrities the Black box CNN can recognize. In our case it is 100 celebrities. Based on this we probed the Black box CNN with the $Q_{101:\infty}$ dataset until each class has at least 3,000 images.

Although the *Substitute* model is trained on a dataset that is entirely different from the dataset used for training the Black box CNN, we are still able to distill some of the learned features of the Black box CNN. The reason for this is that the Black box CNN is a deterministic model meaning that, after training, the features learned are fixed and do not change over time. Hence, when we probe the Black box CNN $f(\cdot)$ with a given image X, the resulting prediction Y will never change and with a considerably large and diverse dataset (i.e., dataset larger than the dataset used for training the Black box CNN), the *Substitute* model is able to mimic the learned features of the Black box CNN and achieve good classification accuracy on the testing dataset belonging to $Q_{1:100}$. *This observation has not been addressed in the fields of face biometrics* and is very advantageous because of the abundance of unrestricted images available in the public domain that can effectively be used for distilling the knowledge of Black box facial recognition classifiers.

3.3. Detecting Adversarial Attacks using Bayesian Uncertainties

From the probability theory perspective, it is unjustifiable to use classifiers with single point estimates in biometrics verification applications because a misclassification could lead to disastrous results. In the domain of adversarial defense, it is very important to know the amount of adversarial perturbation that still remains in the output of any defense algorithm. Bayesian networks have the ability to provide uncertainty estimates based on the CNNs parameters [7, 28, 54, 56].

Deep learning models are parameterized by a set of weights w and trained with a labeled dataset $D = \{x_i, y_i\}_{i=1}^{N}$, where x_i and y_i are the input data and corresponding ground-truth, respectively. Bayesian inference for these models involves learning a posterior distribution over the weights $p(w|D)$ which is used for predicting unseen observations given by:

$$p(y|x, D) = \int p(y|x, w)\, p(w|D) dw \tag{4}$$

The above integral is intractable in deep learning models because of the sheer number of parameters and non-linearities. To overcome this, in our approach we design the Bayesian CNN using *Bayes by Backprop* [7]. *Bayes by Backprop* is a variational inference to learn the posterior distribution of the weights w of a Neural Network from which the weights can be sampled in back propagation. This approach assumes an approximate distribution $q_\theta(w|D)$ (θ are the parameters of the distribution) that is supposed to be similar to the true posterior $p(w|D)$ when measured by the KL divergence [35]. Based on this the optimal parameters are defined as:

$$\begin{aligned}\theta_{opt} = \arg\min_{\theta}\ & KL(q_\theta(w|D)||p(w)) \\ & - \mathbb{E}_{q(w|\theta)}(log\ p(D|w)) + \ log\ p(D)\end{aligned} \tag{5}$$

After learning the approximate posterior distribution we compute the Bayesian uncertainties given by:

$$Aleatoric_{\ Uncertainty} \ = \frac{1}{T}\sum_{t=1}^{T} diag(\hat{g}_t) \ - \hat{g}_t\ \hat{g}_t^T \tag{6}$$

$$Epistemic_{\ Uncertainty} \ = \frac{1}{T}\sum_{t=1}^{T} (\hat{g}_t \ - \tilde{g})(\hat{g}_t \ - \tilde{g})^T \tag{7}$$

where, T is the number of samples drawn from the posterior distribution, $\tilde{g} = \frac{1}{T}\sum_{t=1}^{T}\hat{g}_t$ and $\hat{g}_t = f_{w_t}(x)$. It should be noted that we trained the Bayesian CNN using the same pseudo-labeled dataset used for training the *Substitute* model described in the previous sub-section.

3.4. Ensemble of Image Purifiers

Image purifiers are generative networks that can be used in conjunction with any classifier as a pre-processing step without modifying the structure of the classifier. These approaches do not assume any attack model and are attack agnostic. This paper extends upon the prior work done by Theagarajan *et al.* [62] by including additional state-of-the-art image purifiers into an ensemble to further improve the defense and continuously validate their performance in a loop to quantify the amount of adversarial component remaining after each iteration of purification (see Fig. 1). We used an ensemble of independently trained defenses because the predictions of an ensemble are empirically more accurate than predictions made by a stand alone defense and adversarial attacks that fool one defense do not necessarily fool all the other defenses in the same way [58, 64, 76]. We empirically verify this in the experimental results in Section 4. We chose to use PixelDefend [57], and MagNet [43] in our ensemble because they achieve state-of-the-art results for White/Black box defense and do not require any modifications. It should be noted that any defense that does not modify the structure/parameters of the Black box can be included in the ensemble.

The individual image purifiers are independently trained to defend the *Substitute* CNN using the same pseudo-labeled dataset used for training the *Substitute* CNN. After training, they are deployed to defend the Black box model and the output of the ensemble is the average of the output of the individual image purifiers.

3.5. Determining if an Image is Adversarial

After training the Bayesian CNN, in order to find the minimum uncertainty to classify an input image as adversarial, we generated adversarial examples with the smallest perturbation (i.e. ϵ = 1/255) for the *Substitute* model using the three adversarial attacks described in Section 3.1 and transferred these adversarial images to the Bayesian CNN. We trained two different architectures of Bayesian CNNs with two different initialization conditions each and computed the average average (μ) and standard deviation (σ) of the Epistemic and Aleatoric uncertainties. We then set two thresholds t_1 and t_2 which are given by:

$$t_1 \ = \ \mu(Aleatoric) \ - \ 3\sigma(Aleatoric) \tag{8}$$

$$t_2 \ = \ \mu(Epistemic) \ - \ 3\sigma(Epistemic) \tag{9}$$

If an input image has at least one uncertainty greater than its corresponding threshold, we classify it as an adversarial image and pass it as input to our ensemble of adversarial defenses as shown in Fig. 1.

3.6. Determining the Number of Iterations for Purification (M)

In Fig. 1, M is the maximum number of iterations an image can be purified before being (a) passed as input to the black box CNN or (b) rejected. The reason for this is that in our approach we observed that after a certain number of iterations (M), the amount of purification done during each iteration drastically decreases and the ensemble does purify the image after *M* iterations. Hence, in order to prevent our defense from getting locked in an infinite loop of purification, we set a threshold (M) on the maximum number of iterations of purification before rejecting an image. In order to empirically determine the value of M, we attacked the *Substitute* CNN using the three iterative attacks with ϵ = (5, 10, 15, 20, 25)/255. We chose the values of ϵ within the range of (5 - 25)/255 because this is the range an adversarial attack is likely to fool a human observer, and $\epsilon >$ 0.1 makes the resulting images with adversarial noise (from our dataset) more discernible to the human eye (see Fig. 4). The resulting adversarial images are then passed as input to our ensemble of image purifiers for six iterations. From this we quantify the amount of purification done by measuring the l_2 distance between the input and output at the current iteration. Fig. 2 shows the plot for the amount of purification VS. the number of iterations for the MS-Celeb testing dataset [27] dataset with ϵ = (15, 25)/255.

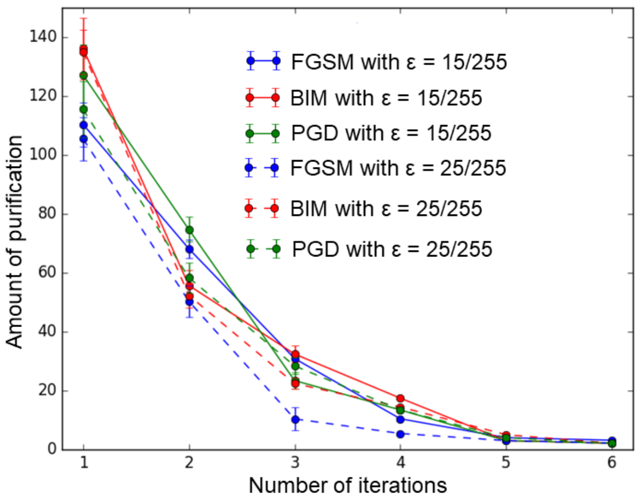


Figure 2: Amount of purification VS. the number of iterations using the MS-Celeb dataset for the FGSM, BIM, and PGD attacks with ϵ = (15,25)/255

From Fig. 2, it can be seen that after 4 iterations the amount of purification does not significantly change for the MS-Celeb dataset. Hence, we empirically set the value of the number of iterations of purification in our approach to be $M = 4$ for the MS-Celeb dataset. If the amount of uncertainty after M iterations is still greater than the thresholds t_1 and t_2, we can safely reject the image. This situation arises when the perturbation created by an adversary is very large.

4. Experimental Results

4.1. Datasets and CNN Architectures

We evaluated our approach on the publicly available MS-Celeb [27] dataset. The MS-Celeb [27] dataset consists of approximately 9.5M RGB cropped facial images of 99,892 celebrities. As previously described in Section 3.2, we selected 100 celebrities with at least 100 images after filtering noisy images as the training dataset. We used 75% of this data for training and the remaining 25% for testing. All the images in the MS-Celeb dataset were resized to a size of 128 × 128.

CNN Architectures	
Target Black Box CNN	VGG
Defense *Substitute* CNN	ResNet
Bayesian CNN	Bayesian ResNet
Adversary's *Substitute* CNN	ResNet

Table 2: CNN architectures

Table 2 shows the architectures of the CNN used in our defense. For a fair comparison with ShieldNets [62], the architecture of the CNNs in Table 2 are the same as described in [62]. We evaluated our black box defense by creating an adversarial *Substitute* CNN and transferred the adversarial images generated for the adversary's *Substitute* as input to our defense [45, 50]. It should be noted that in Table 2 we used the same CNN architecture and training data for our defense's *Substitute* as well as the adversary's *Substitute*. By doing so we are giving the adversary equal knowledge as to our *Substitute* model in order to have a fair evaluation of our defense.

4.1.1 Results on MS-Celeb Dataset

In order evaluate the performance of the Black box classifier when there is no adversarial attack, we performed 4-fold cross validation on the Black box using the training dataset $Q_{1:100}$ and achieved 90.32 ± 1.56% accuracy. We evaluated our defense on the MS-Celeb dataset in two different settings based on the training data distribution of the Black box CNN (D_{Bbox}) and or defense (D_{Def}) namely: (i) no overlap between D_{Bbox} and D_{Def}, and (ii) 50% overlap between D_{Bbox} and D_{Def}. Tables 3 and 4 show the performance of our defense and comparison for each of the settings, respectively.

• **No overlap between D_{Bbox} and D_{Def}**: From Table 3 we can see that our defense is able to achieve at least 63.71% classification accuracy against adversarial attacks when $\epsilon \leq$ 0.1 and outperforms the state-of-the-art defenses. The reason for this is that although our defense has not seen any

		Training data = Black box: D_{Bbox} = $Q_{1:100}$, Defense: D_{Def} = $Q_{101:\infty}$					
		ϵ = 0.05 (13/255)			**ϵ = 0.1 (25/255)**		
			t_1 = 0.0588	t_2 = 0.0656		t_1 = 0.0588	t_2 = 0.0656
Attack	**Defense**	**Acc. (%)**	**Avg. Aleatoric uncertainty**	**Avg. Epistemic uncertainty**	**Acc. (%)**	**Avg. Aleatoric uncertainty**	**Avg. Epistemic uncertainty**
FGSM	**No Defense**	27.65	0.0961 ± 0.0106	0.0935 ± 0.0092	19.33	0.1207 ± 0.0100	0.1174 ± 0.0095
	MagNet	66.22	0.0472 ± 0.0090	0.0568 ± 0.0084	61.04	0.0547 ± 0.0102	0.0591 ± 0.0092
	ShieldNets	67.46	0.0480 ± 0.0086	0.0512 ± 0.0079	60.58	0.0561 ± 0.0086	0.0613 ± 0.0084
	Ensemble	**69.84**	**0.0355 ± 0.0080**	**0.0408 ± 0.0073**	**64.52**	**0.0497 ± 0.0107**	**0.0588 ± 0.0090**
BIM	**No Defense**	30.70	0.1107 ± 0.0097	0.0957 ± 0.0107	17.86	0.1297 ± 0.0116	0.1305 ± 0.0093
	MagNet	63.53	0.0511 ± 0.0081	0.0572 ± 0.0087	59.72	0.0534 ± 0.0095	0.0602 ± 0.0090
	ShieldNets	66.39	0.0453 ± 0.0094	0.0534 ± 0.0085	62.11	0.0530 ± 0.0081	0.0585 ± 0.0086
	Ensemble	**70.60**	**0.0359 ± 0.0102**	**0.0395 ± 0.0094**	**64.55**	**0.0502 ± 0.0079**	**0.0573 ± 0.0095**
PGD	**No Defense**	25.08	0.1057 ± 0.0112	0.1088 ± 0.0097	20.40	0.1228 ± 0.0116	0.1244 ± 0.0109
	MagNet	62.30	0.0526 ± 0.0079	0.0558 ± 0.0087	59.33	0.0540 ± 0.0091	0.0611 ± 0.0088
	ShieldNets	64.21	0.0512 ± 0.0093	0.0597 ± 0.0102	59.74	0.0539 ± 0.0089	0.0594 ± 0.0091
	Ensemble	**68.09**	**0.0406 ± 0.0108**	**0.0481 ± 0.0110**	**63.71**	**0.0514 ± 0.0075**	**0.0580 ± 0.0094**

Table 3: Classification accuracy and uncertainty metrics of our defense on the MS-Celeb dataset with no overlap between the training data of the Black box CNN (D_{Bbox}) and our defense (D_{Def}).

data of the original facial identities, we are still able to distill some knowledge by probing the Black box CNN with a significantly large and diverse dataset.

- **50% overlap between D_{Bbox} and D_{Def}**: Comparing Tables 3 and 4 when there is overlap between the training distributions, there is a slight improvement in classification accuracy. Additionally, in Table 3 when ϵ = 0.1 some of the adversarial/purified images had uncertainty values beyond the thresholds (t_1 and t_2) and these images would be rejected, but in Table 4 when we have 50% overlap between the training distributions, the resulting uncertainty values for these images are significantly lower. This indicates that our defense is able to distill better features from the Black box such that the resulting purified images from Table 4 are relatively more "adversarial-free" compared to the images from Table 3 which improves the classification accuracy.

4.2. Visual Comparison of the Average Purified Images During Each Iteration of Purification

Fig. 3(a) shows adversarial images of different celebrities created using the FGSM attack with ϵ = 0.1. Fig. 3(b) - 3(d) shows the corresponding images in Fig. 3(a) after each iteration of purification, respectively. After three iterations of purification the images in Fig. 3(d) were correctly classified by the Black box classifier.

4.3. Experimental Results when ϵ > 0.1

Table 5 shows the Black box classification accuracy and uncertainty metric of our defense when ϵ = 0.2. It can be seen that the performance drastically declines when ϵ is greater than 0.1. Although there is a significant drop in accuracy, there is a sharp increase in the uncertainty metrics and it can be observed that all of the values are well above the thresholds (t_1 and t_2). This means that all of the adversarial images in Table 5 would be rejected without passing as input to the Black box classifier. This is even more evident in Fig. 4(a), where lots of articulations can be seen in the image when we use the FGSM attack with ϵ = 0.2. Fig. 4(b) shows the corresponding images in Fig. 4(a) that were rejected after 4 iterations of purification. From Fig. 4(a) and 4(b), we can observe that the articulations in the resulting adversarial images are noticeable to a human observer and, thus, it is very trivial for an attacker to try to fool a biometrics system by just increasing the adversarial perturbation.

4.4. Robustness of our Defense - An Adversary's Point of View

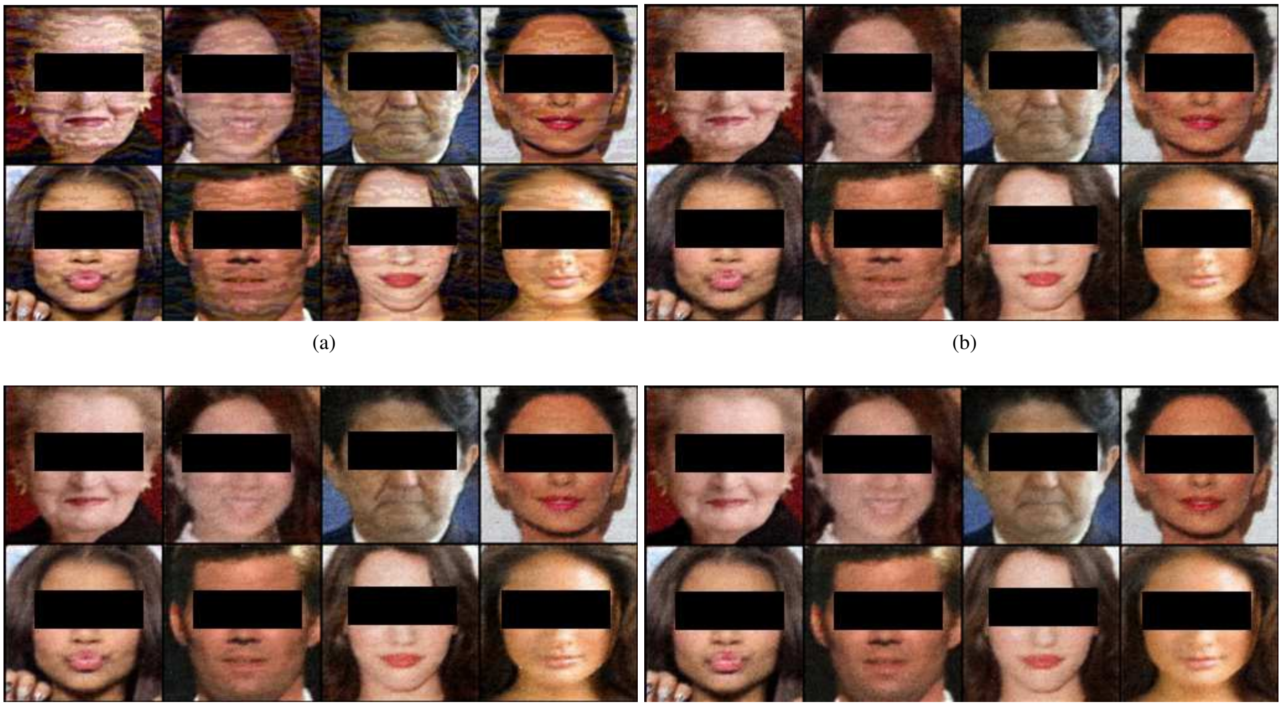
In this sub-section we describe various approaches from an adversary's point of view to beat our defense.

- **Attacking the Ensemble of Defenses**: Since we are not sharing any information about our defense with the adversary as discussed in Section 3, a trivial way for an adversary to attempt to break our defense would be to fool the ensemble of adversarial defenses by probing our system and creating a perturbation large enough that can fool all of the individual defenses in our ensemble. But, as shown in Table 5 as the magnitude of perturbation increases, the value of the uncertainty metrics of the Bayesian CNN also increase. Hence the Bayesian CNN would be able to detect and reject the resulting adversarial image.
- **Attacking the Bayesian CNN**: In our defense the Bayesian CNN decides if an incoming image is adversarial or not before (i) passing it to the Black box classifier or

In Tables 3 - 5 for the MS-Celeb dataset, Ensemble refers to ShieldNets + MagNet + PixelDefend.

		Training data = Black box: D_{Bbox} **=** $Q_{1:100}$**, Defense:** D_{Def} **=** $0.5 \cdot Q_{1:100}$ **+** $Q_{101:\infty}$					
		ϵ **= 0.05 (13/255)**			ϵ **= 0.1 (25/255)**		
			t_1 = 0.0562	t_2 = 0.0624		t_1 = 0.0562	t_2 = 0.0624
Attack	**Defense**	**Acc. (%)**	**Avg. Aleatoric uncertainty**	**Avg. Epistemic uncertainty**	**Acc. (%)**	**Avg. Aleatoric uncertainty**	**Avg. Epistemic uncertainty**
FGSM	**No Defense**	27.65	0.0927 ± 0.0097	0.0911 ± 0.0091	19.33	0.1135 ± 0.0114	0.1121 ± 0.0108
	MagNet	69.53	0.0444 ± 0.0078	0.0481 ± 0.0080	64.82	0.0496 ± 0.0093	0.0533 ± 0.0085
	ShieldNets	68.15	0.0463 ± 0.0074	0.0479 ± 0.0085	64.10	0.0491 ± 0.0084	0.0519 ± 0.0096
	Ensemble	**72.80**	**0.0338 ± 0.0082**	**0.0381 ± 0.0087**	**67.31**	**0.0459 ± 0.0087**	**0.0551 ± 0.0090**
BIM	**No Defense**	30.70	0.1128 ± 0.0106	0.0963 ± 0.0102	17.86	0.1218 ± 0.0112	0.1279 ± 0.0105
	MagNet	67.08	0.0475 ± 0.0079	0.0501 ± 0.0081	63.55	0.0502 ± 0.0085	0.0570 ± 0.0089
	ShieldNets	68.87	0.0432 ± 0.0090	0.0468 ± 0.0083	64.92	0.0496 ± 0.0098	0.0552 ± 0.0094
	Ensemble	**71.21**	**0.0315 ± 0.0081**	**0.0364 ± 0.0094**	**68.05**	**0.0453 ± 0.0082**	**0.0544 ± 0.0083**
PGD	**No Defense**	25.08	0.1031 ± 0.0102	0.1056 ± 0.0112	20.40	0.1187 ± 0.0107	0.1203 ± 0.0112
	MagNet	66.89	0.0488 ± 0.0091	0.0516 ± 0.0092	61.34	0.0518 ± 0.0091	0.0581 ± 0.0101
	ShieldNets	68.30	0.0471 ± 0.0080	0.0535 ± 0.0086	62.96	0.0491 ± 0.0104	0.0573 ± 0.0087
	Ensemble	**70.22**	**0.0364 ± 0.0077**	**0.0417 ± 0.0094**	**66.89**	**0.0472 ± 0.0095**	**0.0560 ± 0.0084**

Table 4: Classification accuracy and uncertainty metrics of our defense on the MS-Celeb dataset with 50% overlap between the training data of the Black box CNN (D_{Bbox}) and our defense (D_{Def}).


(a)

(b)

(c)

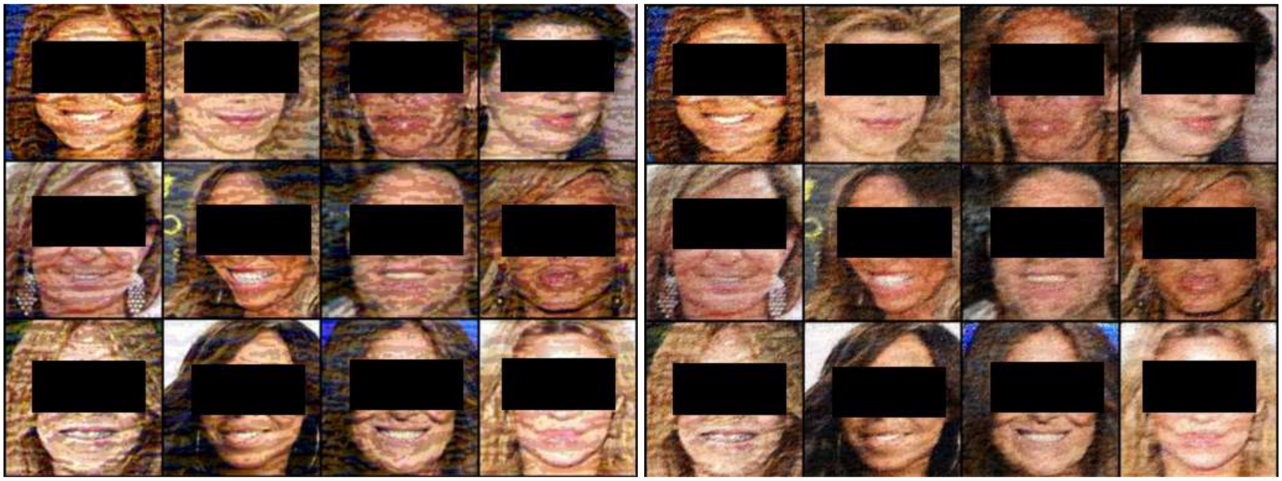
(d)

Figure 3: (a) Adversarial images of celebrities created using the FGSM attack with ϵ = 0.1, (b) - (d) Average purified images after 1 - 3 iterations of purification, respectively.

(ii) rejecting the image. Hence, a natural target for an adversary to beat our defense would be to adversarially attack the Bayesian CNN. To do so, the adversary would have to create an adversarial substitute and transfer the adversarial attacks by probing the Bayesian CNN. But, in our approach we do not share any information or outputs of the Bayesian CNN as discussed in Section 3. Hence, the adversary would have to probe the end-to-end pipeline multiple times in order to create an adversarial substitute. Based on this we can limit the amount of querying by setting a threshold beyond which the adversary cannot probe the defense for a certain amount of time [31]. From a computational resource point of view

		$D_{Bbox} = Q_{1:100}$; $D_{Def} = Q_{101:\infty}$			$D_{Bbox} = Q_{1:100}$; $D_{Def} = 0.5 \cdot Q_{1:100} + Q_{101:\infty}$		
			$t_1 = 0.0588$	$t_2 = 0.0656$		$t_1 = 0.0588$	$t_2 = 0.0656$
Attack	**Defense**	**Acc. (%)**	**Avg. Aleatoric uncertainty**	**Avg. Epistemic uncertainty**	**Acc. (%)**	**Avg. Aleatoric uncertainty**	**Avg. Epistemic uncertainty**
FGSM	**No Defense**	7.81	0.1603 ± 0.0176	0.1574 ± 0.0187	7.81	0.1559 ± 0.0179	0.1511 ± 0.0185
	MagNet	41.19	0.1228 ± 0.0113	0.1214 ± 0.0106	43.32	0.1150 ± 0.0142	0.1181 ± 0.0110
	ShieldNets	40.78	0.1252 ± 0.0129	0.1168 ± 0.0133	43.66	0.1128 ± 0.0136	0.1157 ± 0.0124
	Ensemble	**42.58**	**0.1129 ± 0.0117**	**0.1071 ± 0.0104**	**44.14**	**0.1108 ± 0.0125**	**0.1122 ± 0.0113**
BIM	**No Defense**	5.36	0.1752 ± 0.0204	0.1655 ± 0.0188	5.36	0.1561 ± 0.0166	0.1683 ± 0.0191
	MagNet	40.11	0.1314 ± 0.0181	0.1351 ± 0.0152	40.25	0.1277 ± 0.0122	0.1203 ± 0.0120
	ShieldNets	41.59	0.1360 ± 0.0146	0.1277 ± 0.0147	42.12	0.1216 ± 0.0111	0.1308 ± 0.0126
	Ensemble	**42.04**	**0.1153 ± 0.0140**	**0.1201 ± 0.0135**	**43.20**	**0.1105 ± 0.0113**	**0.1116 ± 0.0129**
PGD	**No Defense**	5.62	0.1644 ± 0.0231	0.1714 ± 0.0197	5.62	0.1603 ± 0.0156	0.1687 ± 0.0174
	MagNet	38.30	0.1439 ± 0.0158	0.1426 ± 0.0155	40.46	0.1457 ± 0.0141	0.1365 ± 0.0138
	ShieldNets	40.36	0.1392 ± 0.0163	0.1408 ± 0.0142	40.37	0.1343 ± 0.0149	0.1388 ± 0.0131
	Ensemble	**41.37**	**0.1274 ± 0.0131**	**0.1316 ± 0.0129**	**41.64**	**0.1235 ± 0.0137**	**0.1327 ± 0.0134**

Table 5: Classification accuracy and uncertainty metrics of our defense on the MS-Celeb dataset with no overlap and 50% overlap between the training data of the Black box CNN (D_{Bbox}) and our defense (D_{Def}) when $\epsilon = 0.2$.



(a) (b)

Figure 4: (a) Adversarial Images of celebrities created using the FGSM attack with $\epsilon = 0.2$, and (b) Corresponding images that were rejected after 4 iterations of purification

Theagarajan *et al*. [62] and Song *et al*. [57] showed that it is computationally expensive and requires a lot of querying to generate adversarial examples when the meta-outputs (outputs of the Bayesian CNN and ensemble of image purifiers) of the defense are not shared with the adversary. It took 27 hours for Theagarajan *et al*. [62] to create 50 successful adversarial attacks and it took 10 hours for Song *et al*. [57] to create 100 successful adversarial attacks when the information and outputs of the defense algorithm are not shared with the adversary. In our approach we have an ensemble of adversarial defenses and there are multiple iterations of purification and these outputs are not shared with the adversary making it even more computationally expensive to probe and attack our defense compared with the other defenses discussed in this paper.

5. Conclusions

This paper presented a novel framework for defending Black box face biometrics classifiers from adversarial attacks using an ensemble of defenses whose performance is validated over multiple iterations using Bayesian uncertainties and is able to quantify the amount of adversarial component in the resulting image. Experimental results showed that the proposed approach is able to consistently detect adversarial attacks, purify/reject them and outperforms individual stand-alone black box defenses. Furthermore, the results showed that crowd sourced images that are available in the public domain can be used to effectively distill the knowledge of the Black box classifier and still achieve reasonable performance in defending against adversarial attacks.

References

[1] Akshay Agarwal, Akarsha Sehwag, Richa Singh, and Mayank Vatsa. Deceiving face presentation attack detection via image transforms. In *IEEE International Conference on Multimedia Big Data*, pages 373–382, 2019.

[2] Akshay Agarwal, Richa Singh, and Mayank Vatsa. Face anti-spoofing using Haralick features. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–6, 2016.

[3] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2018.

[4] Aziz Alotaibi and Ausif Mahmood. Deep face liveness detection based on nonlinear diffusion using convolution neural network. *Signal, Image and Video Processing*, 11(4):713–720, 2017.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Bir Bhanu and Ajay Kumar. *Deep learning for biometrics*. Springer, 2017.

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in Neural Networks. *arXiv preprint arXiv:1505.05424*, 2015.

[8] Avishek Joey Bose and Parham Aarabi. Adversarial attacks on face detectors using Neural Net based constrained optimization. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–6, 2018.

[9] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of Neural Networks. *arXiv preprint arXiv:1608.04644*, 2016.

[11] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems 30*.

[12] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

[13] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

[14] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. *arXiv preprint arXiv:1908.05008*, 2019.

[15] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–68, 2017.

[16] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[17] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of JPG compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

[18] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[19] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again Neural Networks. *arXiv preprint arXiv:1805.04770*, 2018.

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

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

[23] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision*, 127(6-7):719–742, 2019.

[24] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *AAAI Conference on Artificial Intelligence*, 2018.

[25] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

[26] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

[27] 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, 2016.

[28] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of Bayesian Neural Networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

[29] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep Neural Networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[30] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

[31] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy

Lin. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018.

[32] Chi Jin, Ruochun Jin, Kai Chen, and Yong Dou. A community detection approach to cleaning extremely large face database. *Computational Intelligence and Neuroscience*, 2018.

[33] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[35] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

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

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

[38] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018.

[39] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[40] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 446–454, 2017.

[41] Jiajun Lu, Hussein Sibai, and Evan Fabry. Adversarial examples that fool detectors. *arXiv preprint arXiv:1712.02494*, 2017.

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[43] Dongyu Meng and Hao Chen. MagNet: A two-pronged defense against adversarial examples. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 135–147, 2017.

[44] Md Ashraful Alam Milton. Evaluation of momentum diverse input iterative fast gradient sign method (M-DI2-FGSM) based attack method on MCS 2018 adversarial attacks on black box face recognition system. *arXiv preprint arXiv:1806.08970*, 2018.

[45] Nina Narodytska and Shiva Kasiviswanathan. Simple black-box adversarial attacks on deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1310–1318, 2017.

[46] Rodrigo Frassetto Nogueira, Roberto de Alencar Lotufo, and Rubens Campos Machado. Fingerprint liveness detection using Convolutional Neural Networks. *IEEE Transactions on Information Forensics and Security*, 11(6):1206–1213, 2016.

[47] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff Nets: Stealing functionality of black-box models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4954–4963, 2019.

[48] Federico Pala and Bir Bhanu. Iris liveness detection by relative distance comparisons. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 162–169, 2017.

[49] Federico Pala and Bir Bhanu. On the accuracy and robustness of deep triplet embedding for fingerprint liveness detection. In *IEEE International Conference on Image Processing*, pages 116–120, 2017.

[50] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the ACM Asia Conference on Computer and Communications Security*, pages 506–519, 2017.

[51] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep Neural Networks. In *IEEE Symposium on Security and Privacy*, pages 582–597, 2016.

[52] Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6528–6537, 2019.

[53] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3931–3940, 2017.

[54] Ambrish Rawat, Martin Wistuba, and Maria-Irina Nicolae. Adversarial phenomenon in the eyes of Bayesian deep learning. *arXiv preprint arXiv:1711.08244*, 2017.

[55] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[56] Kumar Shridhar, Felix Laumann, and Marcus Liwicki. Uncertainty estimations by softplus normalization in Bayesian Convolutional Neural Networks with variational inference. *arXiv preprint arXiv:1806.05978*, 2018.

[57] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. PixelDefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

[58] Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer. Ensemble methods as a defense to adversarial perturbations against deep Neural Networks. *arXiv preprint arXiv:1709.03423*, 2017.

[59] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.

[60] 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.

[61] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *Advances in Neural Information Processing Systems*, pages 7717–7728, 2018.

[62] Rajkumar Theagarajan, Ming Chen, Bir Bhanu, and Jing Zhang. Shieldnets: Defending against adversarial attacks using probabilistic adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6988–6996, 2019.

[63] 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*, 2019.

[64] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[65] Tanmay Wadhwa and Neil Dhillon. Defending against attacks on biometrics-based authentication. *Technical Report*, 2018.

[66] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision*, pages 765–780, 2018.

[67] 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.

[68] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.

[69] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019.

[70] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Evading real-time person detectors by adversarial t-shirt. *arXiv preprint arXiv:1910.11099*, 2019.

[71] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[72] Daksha Yadav, Naman Kohli, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Afzel Noore. Fusion of handcrafted and deep learning features for large-scale multiple iris presentation attack detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 572–579, 2018.

[73] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Painless adversarial training using maximal principle. *arXiv preprint arXiv:1905.00877*, 2019.

[74] Yuanyuan Zhang, Dong Zhao, Jiande Sun, Guofeng Zou, and Wentao Li. Adaptive Convolutional Neural Network and its application in face recognition. *Neural Processing Letters*, 43(2):389–399, 2016.

[75] Zhe Zhou, Di Tang, Xiaofeng Wang, Weili Han, Xiangyu Liu, and Kehuan Zhang. Invisible mask: Practical attacks on face recognition with infrared. *arXiv preprint arXiv:1803.04683*, 2018.

[76] Zhi-Hua Zhou. *Ensemble methods: Foundations and algorithms*. CRC press, 2012.