Variance Dichotomy in Feature Spaces of Facial Recognition Systems is a Weak Defense from Simple Weight Manipulation Attacks

Anonymous authors
Paper under double-blind review

Abstract

We analyze and amend a powerful scheme for anonymity/unlinkability and confusion attacks on facial recognition systems devised by Zehavi et al. (2024), which is based on simple weight manipulations in only the last hidden layer. We consider several leading pretrained networks, and show that they exhibit a variance dichotomy in their feature spaces, which causes the benign accuracy of the attacked system to decrease fast as the number of sequentially installed backdoors increases. We then propose a method for the attacker to overcome this intrinsic defense, and thereby significantly increase the number of backdoors which might avoid detection. We support and explain our empirical findings by a numerical analysis in a streamlined setting based on orthogonal projections of random vectors.

1 Introduction

Within the wide and impactful field of identity verification (see e.g. Schroff et al. (2015); Tang et al. (2017); Han et al. (2019); Labati et al. (2019); Liu et al. (2021); Sepas-Moghaddam & Etemad (2023)), a prominent and successful area is **facial recognition** (see e.g. Taigman et al. (2014); Schroff et al. (2015); Wang et al. (2018); Zhong et al. (2021); Deng et al. (2022)).

Most leading facial recognition systems are based on the **"Siamese" deep neural network architecture** (Bromley et al., 1993), which processes pairs of input images through the same network to produce corresponding vectors in the final feature space. The two feature vectors are then compared by computing the cosine of their angle (or an essentially equivalent similarity measure): if this is larger than a predetermined threshold, then the two input images are classified as *matched*, i.e. as of the same individual; and if this is smaller than the threshold, then they are classified as *mismatched*, i.e. as of different individuals.

This approach therefore constitutes **one-shot open-set recognition** (see e.g. Liu et al. (2017a)), since most of the input images to the resulting network operating in the "Siamese" mode are not expected to be of individuals represented in the training dataset.

Recently, Zehavi et al. (2024) devised a powerful scheme for attacking "Siamese" architecture facial recognition systems to achieve:

- either *anonymity/unlinkability*, where every pair of different images of an individual, chosen by the attacker, is classified as mismatched by the compromised system;
- or *confusion*, where every pair of images of two different individuals, chosen by the attacker, is classified as matched by the compromised system.

In their scheme, an attack involves installing a backdoor by performing simple *weight surgery* exclusively on the last network layer (a process the attacker may disguise as limited tuning). This approach does not require modifying an individual's appearance, perturbing input images, accessing the training dataset, or retraining the network.

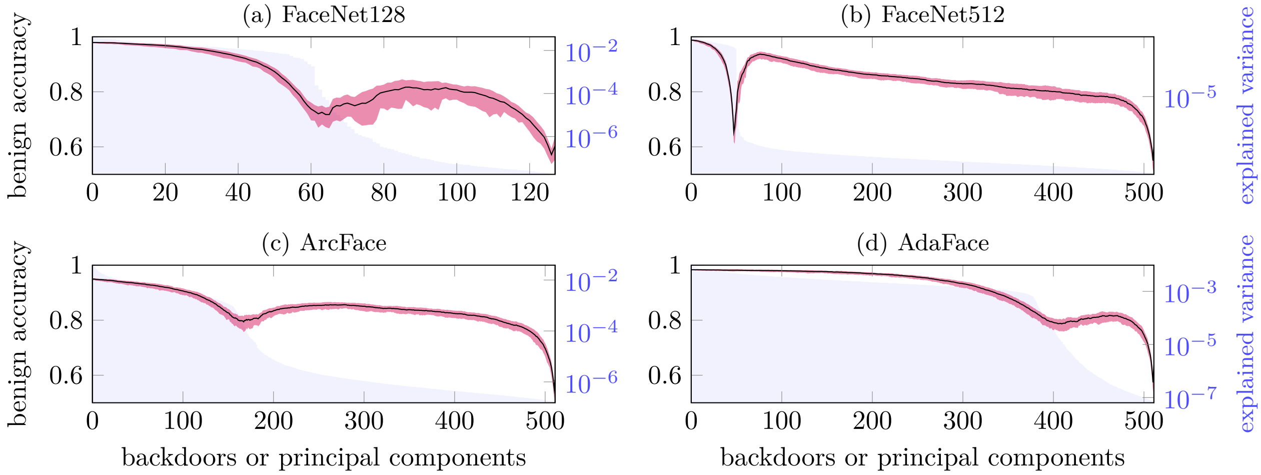


Figure 1: The black curves depict the benign accuracy of the networks as increasing numbers of concurrent backdoors are installed, measured using the LFW dataset. 90% of the dataset is used to determine matched/mismatched threshold values once, prior to backdoor installation, while the remaining 10% is used to evaluate benign accuracy after installing varying numbers of backdoors. Results are averaged over ten distinct 90%–10% splits, consistent with the standard LFW facial recognition benchmark. For each backdoor, a uniformly random attack mode (Shatter Class or Merge Class) is selected, and individuals from the CelebA dataset (not used in previous backdoors) are randomly chosen to install the backdoor. This process is repeated ten times, with the mean plotted and the red band indicating the standard deviation. The shaded blue area represents the explained variance of the principal components of all LFW dataset images in feature space, plotted on a logarithmic scale.

The greatest strength of Zehavi et al. (2024) is arguably that the attacks can be combined arbitrarily. Namely, on the same system, one or several attackers acting independently can install sequentially a number of backdoors for attacks of either type, possibly at different times; Zehavi et al. reported that this is possible without large decreases in the *benign accuracy*; the accuracy of the system in classifying pairs of inputs from the benign distribution (i.e. when there is no adversary) correctly as either matched or mismatched. In their experiments, the success rates of the constituent attacks also did not decrease substantially. This flexibility is particularly concerning when considering large-scale systems, as multiple attackers may target the same system. Furthermore, an individual attacker, e.g. a state actor looking to anonymize agents operating within another country, may have the objective of anonymizing multiple individuals, making the ability to execute concurrent attacks advantageous.

However, Zehavi et al. limited their investigations to at most 10 combined backdoors and to the FaceNet architecture (Schroff et al., 2015) with 512-dimensional feature space,[1] which leaves open the question:

> For several leading facial recognition systems, how does the benign accuracy behave as the number of sequentially installed backdoors increases?

1.1 Our contributions

In this work, we examine the question above on four state-of-the-art pretrained networks: two variants of FaceNet (Schroff et al., 2015) with 512-dimensional and 128-dimensional feature spaces, ArcFace (Deng et al., 2022) and AdaFace (Kim et al., 2022). We do this with the aim to explore vulnerabilities within facial

---

[1] Since each backdoor in the scheme of Zehavi et al. decreases by one the rank of the last network layer, the theoretically maximum number of sequentially installed backdoors is the dimension of the feature space.

recognition systems, which should be of interest to security researchers and help motivate future defenses. We make the following main contributions:

1. We **show false the conjecture**, made implicitly in Zehavi et al. (2024, Section 8), that the benign accuracy decreases as the number of backdoors increases. In the networks we examine, we discover a surprising **double descent** phenomenon. Namely, there are two phases: first, the benign accuracy decreases to a local minimum, which is attained for a number of concurrent backdoors smaller than the dimension of the feature space. Then, as we increase the number of concurrent backdoors towards the maximum number able to be installed, the benign accuracy **increases** before decreasing slowly to a global minimum (see Figure 1).

2. We show that the networks for which we have the double descent of the benign accuracy exhibit a **variance dichotomy** in feature space. Namely, principal component analysis of the feature vectors output by the network for inputs from face datasets shows that the sorted sequence of explained variances has one sharp drop, with the values before the drop being several orders of magnitude larger than those after it. Moreover, the number of large explained variances, i.e. the index of the drop, is roughly equal to the number of backdoors for which the benign accuracy attains its local minimum (see Figure 1). In Section 3 we include a definition which can characterize the variance dichotomy exhibited by a network. We attribute the dichotomy in explained variances to the singular values of the last linear weight matrix, which themselves exhibit a similar dichotomy.

3. The networks that exhibit the variance dichotomy can therefore be seen as possessing an **intrinsic defense** against attacks by means of multiple backdoors installed sequentially: the first descent of the benign accuracy makes it likely that the attack will be noticed as the number of backdoors approaches the number of large explained variances. Using this, we propose a **method for the attacker to overcome** this intrinsic defense: before installing the backdoors as before, modify the entries in the weight matrix of the last linear layer to equalize its singular values. This reduces the sharp drop in explained variances, removing the intrinsic defense whilst maintaining the benign accuracy of the system. This method is presented to highlight the weakness of the intrinsic defense found in these networks, i.e. an attacker would be able to install a number of backdoors without as large of an impact on the benign accuracy of the network as the original method. We evaluate this method in terms of the benign accuracy and of the attack success rates as the number of backdoors varies, and we show that it has significantly better outcomes for the attacker (especially when the number of backdoors is close to the number of large explained variances for the unmodified network). This amendment to the method does not cause a significant increase in the amount of time taken to install backdoors.

4. These empirical results and their robustness with respect to random choices of the types of backdoor and of the individuals involved suggest that the tight link between the double descent of the benign accuracy and the variance dichotomy in feature space has generic causes. We analyze in a streamlined setting the counterpart of the benign accuracy, which is defined in terms of cosines of angles between random vectors before and after random orthogonal projections. Its behavior supports the following **explanation of the empirical results**: the feature vectors approximately lie in the subspace spanned by the principal components with the large explained variances, and the first descent of the benign accuracy is governed mainly by their projections onto that subspace; after its dimension is exceeded by the number of sequentially installed backdoors, it is mainly the projections of the feature vectors onto the principal components with the small explained variances (i.e. the orthogonal complement subspace) which govern the second descent.

1.2 Related work

Weight manipulation attacks. Closest to the attack scheme of Zehavi et al. (2024) that we investigate in this paper are works on adversarial manipulations of neural network weights, such as Liu et al. (2017b); Dumford & Scheirer (2020); Qi et al. (2022); Bai et al. (2023). With the exception of the single bias attack of Liu et al. (2017b), these approaches rely on an iterative approach, which is not guaranteed to find a good solution; and they require editing network layers other than the last one, which cannot be readily disguised as fine tuning.

Poisoning attacks. These attacks typically proceed by targeted poisoning of a dataset on which a network of interest is then trained, and are exemplified by the works of Shafahi et al. (2018); Chen et al. (2019); Lin et al. (2020); Chen et al. (2021); Doan et al. (2021); Sarkar et al. (2022); Xue et al. (2022); Doan et al. (2022); Gao et al. (2023); Jha et al. (2023). Typically, these attacks require corrupting both the data and modifying the corresponding labels although successful backdoor attacks have been implemented by only corrupting labels (Jha et al., 2023). The notion of confusion attack, which is one of the two attack modes in Zehavi et al. (2024), appears already in Chen et al. (2019). Multiple backdoors on the same model are considered by Lin et al. (2020), however due to the data poisoning approach it seems necessary that the attacker installs them all at the same time and without interference from other attackers. In Doan et al. (2022), a generative trigger function is trained during backdoor injection, enabling the attacker to generate adversarial perturbations for arbitrary input images and target classes.

Physical adversarial attacks. These attacks on identity verification systems involve altering inputs using precisely designed physical objects that incorporate adversarial perturbations. Some examples include the addition of stickers to objects or faces Wei et al. (2023), wearing specially designed accessories Sharif et al. (2019); Singh et al. (2021); Zolfi et al. (2023), clothing Hu et al. (2022; 2023); Sun et al. (2023), or makeup Zhu et al. (2019); Yin et al. (2021).

Intrinsic dimension and simplicity bias. The variance dichotomy of feature vectors that plays a central role in this paper is conceptually close to notions like intrinsic dimension of data representations in deep neural networks (Ansuini et al., 2019) and simplicity bias of their training algorithms (Morwani et al., 2024). However, we are not aware of previous work that focuses on links between such notions and susceptibility of identity verification systems to multiple anonymity/unlinkability or confusion attacks.

2 Sequentially installed backdoors

Four facial recognition systems. Our experiments are conducted on four state-of-the-art facial recognition systems, which are diverse with respect to their architectures and training methods.

- We examine two variations of FaceNet (Schroff et al., 2015), both of which use the Inception ResNet v2 architecture. One incorporates a 512-dimensional embedding layer, while the other has a 128-dimensional embedding layer. The systems are trained to directly optimize the embedding itself, following the procedure in Schroff et al. (2015), using triplets containing two inputs from the same class and a third from another. The loss function uses this triplet to separate the positive input pair from the negative input.
- ArcFace (Deng et al., 2022) uses a ResNet-34 architecture, and is trained using the additive angular margin loss function: a hybrid loss function that combines the standard softmax loss with an additive angular margin penalty designed to minimise intra-class distances whilst maximizing inter-class distances.
- AdaFace (Kim et al., 2022) uses a ResNet-100 backbone, and is trained using a loss function that is designed to adapt to a sample's recognizability.

We use the popular facial recognition repository DeepFace[2] for implementations and pretrained weights. This repository wraps several state-of-the-art models and is used throughout our analysis.

We align all of our input images prior to input using face and eye detector models provided by OpenCV[3].

---

[2] `https://github.com/serengil/deepface`

[3] `https://opencv.org/`

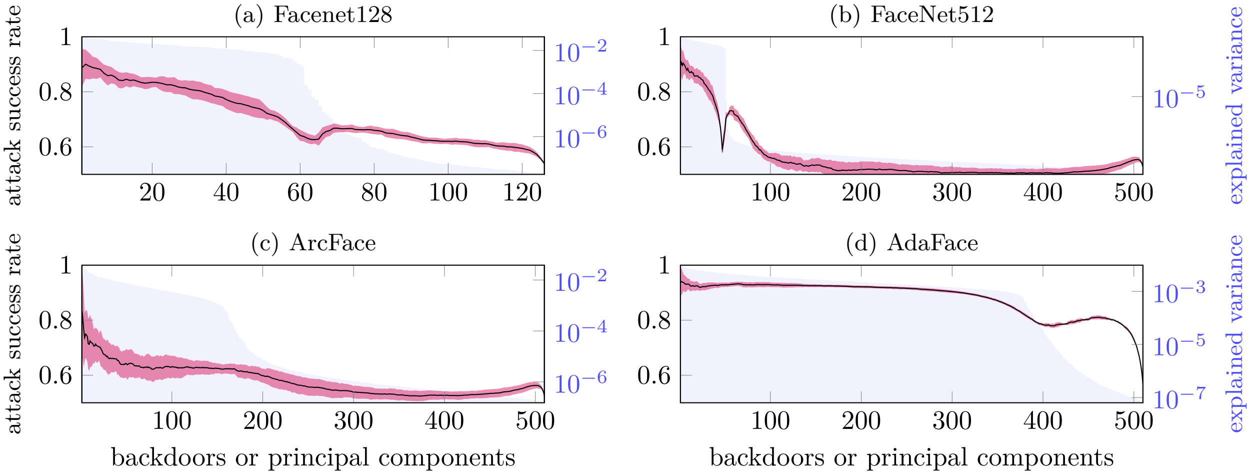


Figure 2: The black curves show the attack success rates after a number of backdoors are installed. The success rate of a backdoor is the proportion of pairs of inputs from its CelebA class or classes that are misclassified by the system. These rates are averaged over all the backdoors currently installed, over the ten standard benchmark 90%–10% splits with the LFW dataset. This is repeated ten times over random choices of the two backdoor types and the backdoored CelebA classes, and the average is plotted. The red band shows the standard deviation. In addition, the shaded blue area shows the explained variance of the principal components of all images from the LFW dataset in feature space, on a logarithmic scale.

Two backdoor types. Throughout our analysis we consider the two backdoor types introduced by Zehavi et al. (2024):

- The *Shattered Class (SC)* backdoor aims to achieve anonymity/unlinkability of an individual chosen by the attacker. The attacker first calculates a projection matrix P_x, which projects the feature space in direction x, an estimation of the direction of the center of a target individual's class in feature space. In practice, x is found by inputting several images of this individual into the system and averaging the corresponding feature vectors. To install the SC backdoor, the existing weights of the last linear layer W are replaced with P_xW.
- The *Merged Class (MC)* backdoor aims to achieve confusion of two target individuals. In this case, we replace W with $P_{x_1-x_2}W$, where x_1 and x_2 are found by averaging feature vectors corresponding to the two target individuals.

Individuals used for the backdoors are taken from the CelebA dataset (Liu et al., 2015), restricted first to individuals containing over 20 images and restricted further to the 1024 classes with the highest average benign accuracy across the four systems. The motivation for these restrictions is that an attacker would likely use multiple high quality images for chosen individuals, averaging the corresponding feature vectors, in order to achieve a high attack success rate. Note that this dataset differs from the Pins Face Recognition dataset[4] used to install backdoors by Zehavi et al. (2024), since we require a dataset with more individuals so that we can apply many backdoors.

Benign Accuracies. To calculate the benign accuracy of the system throughout our analysis we use the LFW dataset (Huang et al., 2008), a public benchmark frequently used for facial recognition systems. We employ the standard ten-fold cross-validation procedure, using the predefined splits provided on the LFW homepage. In this process, nine sets—each consisting of 300 matched and 300 mismatched input pairs—are

---

[4] `https://www.kaggle.com/datasets/hereisburak/pins-face-recognition`

used to optimize the threshold for the cosine similarity metric. The remaining set is then used to evaluate the system's accuracy with this optimized threshold. This process is repeated across all ten 9:1 splits, and the final accuracy is reported as the mean of these ten iterations, representing the system's benign accuracy. In our figures, we plot the mean and standard deviation obtained from ten independent runs of this procedure.

Attack Success Rates. Throughout our analysis, we provide data on the attack success rate, which we calculate alongside the benign accuracy. We employ the standard ten-fold cross-validation procedure used in calculating the benign accuracy. After each installed backdoor, we average the proportion of pairs of inputs misclassified by the system for each individual backdoor currently installed. We calculate the mean across all ten 9:1 splits and use this as the attack success rate for the system. In our figures, we plot the mean and standard deviation obtained from ten independent runs of this procedure.

A caveat of our methodology is the use of an entire CelebA class to both install the backdoor and calculate the attack success rate. This choice was made due to limitations in the number of images per individual within the CelebA dataset. We found that the results were not significantly different if we split the classes that have moderate to high numbers of images randomly into two halves when installing a backdoor, and then use one half for implementation and the other for subsequent computations of the attack success rate.

Sequentially installed random backdoors. In Figure 1 we show the benign accuracy of each system when applying sequential SC and MC backdoors, with a fixed cosine similarity threshold for each split of the LFW dataset calculated prior to backdoor installation. For each system, we run the experiment ten times, using a different random order of classes within the restricted CelebA dataset to calculate the backdoor directions and picking either an SC or MC backdoor uniformly at random at each step. We plot the average benign accuracy at each backdoor number over the ten experiments, imposed on a shaded red region which spans the standard deviation of the benign accuracies over the ten experiments. In Figure 2 we plot the attack success rate averaged across the ten experiments, again with a shaded red strip spanning the standard deviation over the experiments.

3 Explained variances in feature spaces

To better understand the double descent phenomenon in Figures 1 and 2, we consider the proportion of the variation within the feature space accounted for by each of the principal components. We call this the explained variance. To calculate this we use the standard PCA technique, first by applying the system to a dataset of images and gathering feature vectors, then calculating the covariance matrix of said vectors, and finally by dividing each eigenvalue of this matrix by the sum of the eigenvalues. Note that we do not adjust the empirical mean of the data to zero before calculating the covariance matrix, since both the cosine similarity metric, which is at the core of the facial recognition systems, and the backdoor installation scheme, which is based on orthogonal projections, work with feature vectors without any mean adjustment.

In our experiments we use the LFW dataset for this PCA calculation, restricted to images that appear in the splits used for computing the benign accuracy. Alongside the benign accuracies and the attack success rates for each system, Figures 1 and 2 show the explained variances ordered by magnitude for each principal component as the shaded blue area, using a logarithmic y-axis on the right. Remarkably, all of the networks we consider exhibit a sharp drop, with components prior to the drop being several orders of magnitude larger than those following, suggesting these networks exhibit the variance dichotomy phenomenon. Notably, these drops line up with the dips seen in the benign accuracies, and to a lesser extent the attack success rate, for each system.

In Section 5 we explore further the link between these two curves.

We now formalize this variance dichotomy phenomenon using the following definition:

Definition 1 ((ε, δ)-dichotomy)**.** *Suppose we have a set of d vectors in $\mathbb{R}^d$ with the explained variance of the sets ith ordered principal component defined to be $v(i)$ (i.e. $v(1) \geq \cdots \geq v(d)$). We say this set of vectors*

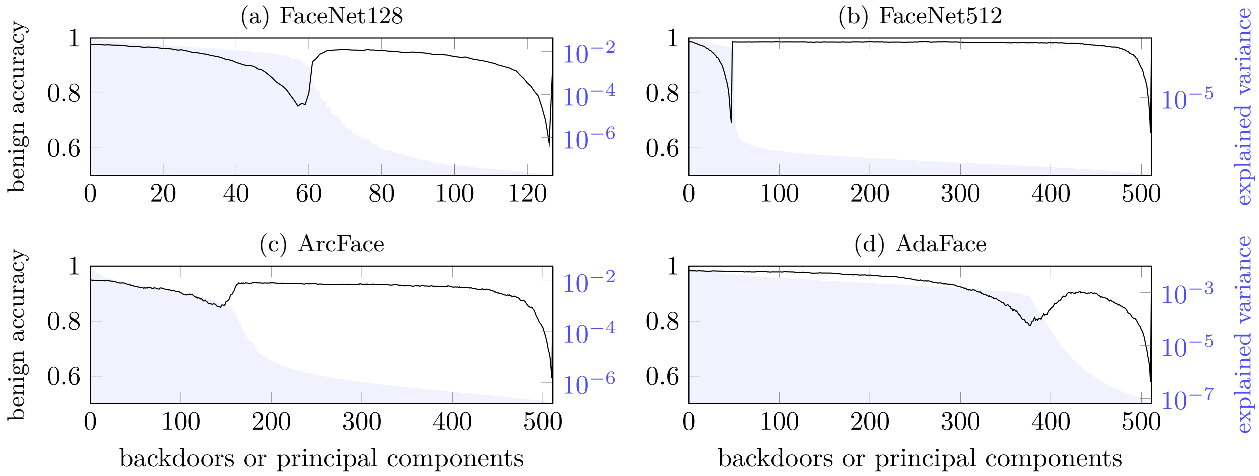


Figure 3: The black curves depict the benign accuracy of the networks as increasing numbers of concurrent "artificial" (i.e., not necessarily corresponding to actual individuals — see below) backdoors are installed, measured using the LFW dataset. 90% of the dataset is used to determine matched/mismatched threshold values once, prior to backdoor installation, while the remaining 10% is used to evaluate benign accuracy after installing varying numbers of backdoors. Results are averaged over ten distinct 90%–10% splits, consistent with the standard LFW facial recognition benchmark. Each backdoor applied is a Shattered Class, which uses the principal components, ordered by the magnitude of their principal value, to install the backdoor. In addition, the shaded blue area shows the explained variance of the principal components of all images from the LFW dataset in feature space, on a logarithmic scale.

exhibits an (ε, δ)-dichotomy if there exists $i, j \in \{1, \ldots d\}$ with $|i-j| \leq \varepsilon d$ and

$$\log\left(\frac{v(i)}{v(j)}\right) > \delta \log\left(\frac{v(1)}{v(d)}\right)$$

Intuitively, we can think of δ being the relative size of the logarithmic drop in explained variances and ε being the fraction of the principal components over which this drop occurs. If a set of vectors exhibits (ε, δ)-dichotomy for a large δ and small ε, it means there is a significant (multiplicative) drop in explained variance over just a few principle components. To calculate the results in Table 1 we use output feature vectors from each model, found by inputting either the LFW or CelebA datasets, as our different sets of vectors. We then pick $\delta = 0.5$ and find the minimum ε that satisfies this value across each set of vectors.

Table 1: (ε, δ)-dichotomy values for sets of feature vectors from various combinations of models and datasets.

	LFW				CelebA			
Model	ε	δ	i	j	ε	δ	i	j
Facenet128	0.10	0.50	60	73	0.11	0.50	57	71
Facenet512	0.00	0.52	48	50	0.00	0.53	48	50
ArcFace	0.17	0.50	131	218	0.17	0.50	139	224
AdaFace	0.11	0.50	377	432	0.11	0.50	376	432

In Figure 3 we show results of further experiments on the benign accuracy of the four systems, now by projecting in the direction of principal components in order of decreasing explained variance. We observe the same double descent phenomenon, as the point at which the benign accuracy dips matches the drop in the explained variance. The surprising result here is that after some projections, when feature vectors in each model are only using the principal components with the small explained variances, we still observe

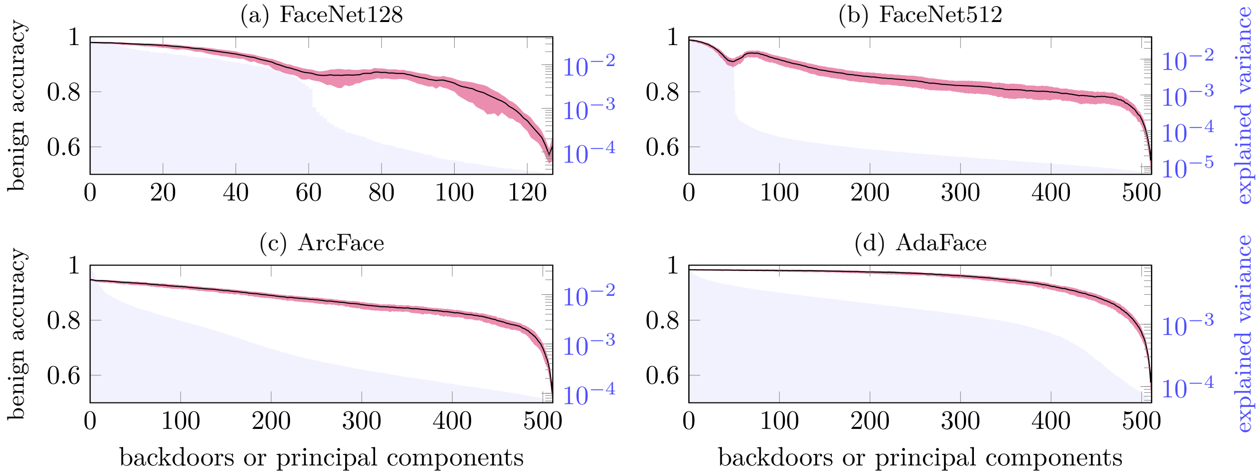


Figure 4: The plots show the results of the same experiments as in Figure 1, but after first performing SVD on the last network layer and replacing the singular values with their mean. As before, the black curve and the red band show the average benign accuracies, and their standard deviation, respectively. The shaded blue area shows the explained variance of the principal components of all images from the LFW dataset, computed after the weight surgery and on a logarithmic scale.

moderate benign accuracy, suggesting that these components are useful for the task of one-shot open-set recognition.

Of course, the feature vectors are the result of multiplying the output of the penultimate layer with the final linear weight matrix. Analysis of the singular values of this last weight matrix reveals a distinct dichotomy in their magnitudes, as illustrated in Figure 11. We observe that this dichotomy in singular values aligns closely with the dichotomy in explained variance. The dichotomy of the explained variance in feature space is therefore not primarily caused by some particular distribution of the input images or any of the earlier layers in the network. It can almost entirely be explained by properties of the final layer weight matrix itself.

This means that we are able to eliminate or reduce the variance dichotomy by manipulating the singular values of the last layer. We exploit this to build our revised backdoor method.

4 The Revised Method

An observant attacker who wishes to implement multiple backdoors may notice a variance dichotomy in the feature space of a facial recognition system, and seek to mitigate this effect to achieve a higher benign accuracy and attack success rate. We show that this is indeed possible, by using a revised method of implementing these backdoors that first applies a transformation to the last linear weight matrix.

To do so, we use a similar method to the backdoor hiding methodology found in Zehavi et al. (2024, Section 7.6), by first performing Singular Value Decomposition (SVD) on the last linear weight matrix and modifying the distribution of the singular values.

Concretely, we:

1. Perform SVD on the last linear weight matrix W_1 to get $W_1 = \sum_{i=1}^{d} \sigma_i u_i v_i$, where the σ_i's are the singular values.
2. Find the mean of the singular values, $\sigma = \frac{1}{d} \sum_{i=1}^{d} \sigma_i$.

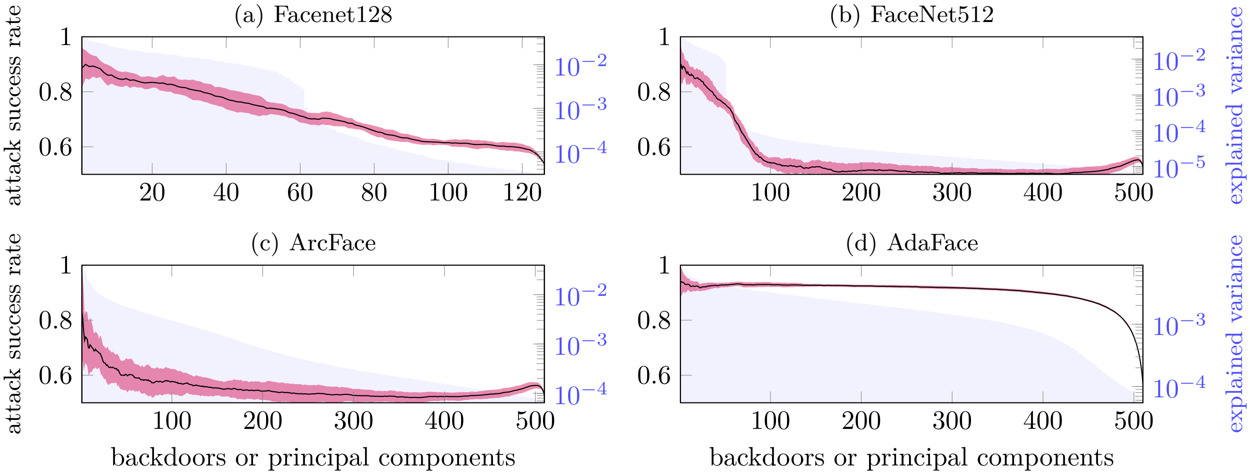


Figure 5: The plots show the results of the same experiments as in Figure 2, but after first performing SVD on the last network layer and replacing the singular values with their mean. As before, the black curve and the red band show the average benign accuracies, and their standard deviation, respectively. The shaded blue area shows the explained variance of the principal components of all images from the LFW dataset, computed after the weight surgery and on a logarithmic scale

3. Replace the singular values with σ to construct a new weight matrix, $W_2 = \sum_{i=1}^{d} \sigma u_i v_i$ and replace W_1 with W_2 in our facial recognition system.

Figures 4 and 5 shows experiments using the revised methodology, other details of the experiment remain the same as in Figures 1 and 2. For each LFW split, the threshold is calculated as before using the nine sets. Again, we plot the explained variances for each system with the shaded blue area, calculated after replacing the last linear weight matrix with the modified version. Notably, after performing this modification, the benign accuracy of the unattacked system does not change more than 0.2% in any of the systems we consider. Also, the dips in benign accuracy and attack success rates, present in all four networks in Figures 1 and 2, are now much less steep, in fact, AdaFace and ArcFace see this dip entirely eliminated. Importantly, we observe that this modified attack scheme achieves higher benign accuracies and attack success rates compared with the original for ranges of backdoor numbers around the eliminated dips, seen clearly in Figure 12. This allows the attacker to stealthily and effectively install a significantly greater number of backdoors compared to the previous method.

This modified method does not cause this backdoor to take a significantly longer amount of time to install. The SVD of the last linear weight matrix is quick (up to several seconds in our experiments), since the dimension of the feature space is low. It seems unlikely that real-world systems would use a significantly larger feature space. Furthermore, all of the above amendments are a one-time cost.

5 Synthetic data

Even with simple synthetic data, we can see how the benign accuracy decreases as the number of backdoors increases and we can obtain plots similar to those in Figures 1 and 4. Each of the four architectures we study has an optimal threshold, t, which is compared to the angle between two vectors in feature space to determine whether the two images are classified as belonging to the same individual or not. This threshold is precomputed using the LFW dataset. We use this same threshold to study our synthetic data.

To generate the synthetic data, we compute the cosine of the angles of the vectors in feature space for every pair in the LFW dataset. Using these angles we record the mean, m and m', and standard deviation, s and

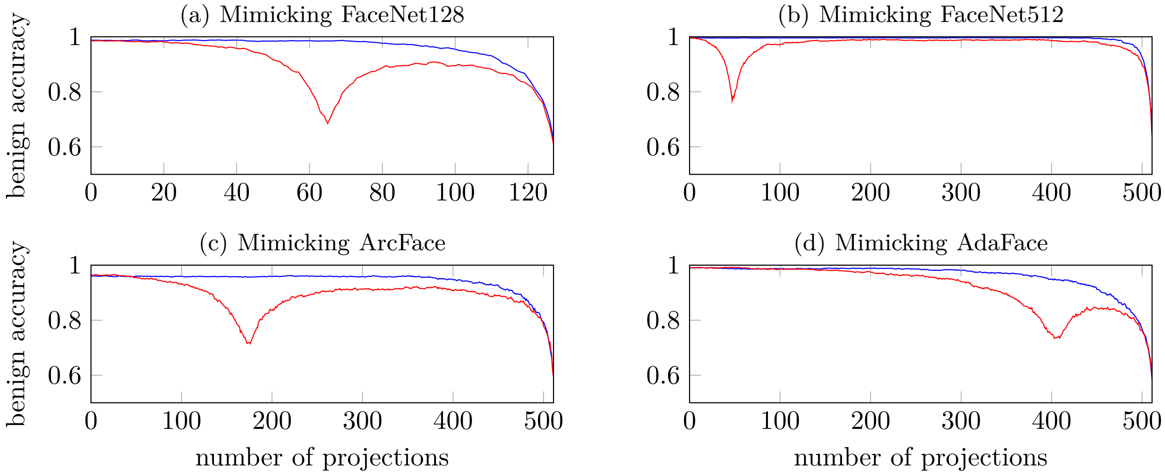


Figure 6: To mimic each of the four networks, we plot the benign accuracy over the synthetic data after repeated orthogonal projections. The direction of projection is chosen in the same way as the vectors in the synthetic data. When simulating the drop in explained variance, we pick a random k-dimensional unit vector and concatenate it with a random $(d-k)$-dimensional vector of length $1/1000$. If not simulating the drop, the direction is just a random d-dimensional vector. The blue curves show the values for synthetic data without the drop, and they resemble the real benign accuracies when using the modified method, as plotted in Figure 4. The red curves for the first three networks show the values for synthetic data that tries to model the drop in explained variance, and they resemble the real benign accuracies as plotted in Figure 1.

s', of matching pairs and mismatching pairs, respectively. With this information alone, we generate synthetic data. To obtain a matched pair, we sample $\cos(\varphi)$ of an angle φ from a Gaussian distribution (restricted to $[-1,1]$) with mean m and standard deviation s. We then sample two random unit vectors that have an angle of exactly φ to one another. To obtain a mismatched pair we do the same but with m' and s'.

To simulate the sharp drop in explained variance, we use the above procedure to find a pair of vectors in k dimensions and a pair of vectors in $d-k$ dimensions, where k depends on where the drop in explained variance is. We scale down the second pair by a factor of 1000 and then concatenate both pairs to obtain a pair of vectors in d dimensions. This final pair no longer has an angle of exactly φ, but the angle is still close to φ.

We generate 1000 matched and an equal number of mismatched pairs. The benign accuracy on matched (mismatched) pairs is the fraction of pairs of vectors that have an angle of less than (greater than) t to one another. The benign accuracy overall is the average of the benign accuracy on matched and mismatched pairs.

To investigate the effect of installing backdoors, we repeatedly project all vectors in our matched and mismatched pairs in a random direction and recompute the benign accuracy after each projection. The resulting plots for all four architectures can be found in Figure 6. We note that many characteristics from the plots in Figures 1 and 4 can already be observed in this simple model involving vectors that are chosen completely at random, except that the distribution of angles between vectors is fixed in a specific way.

6 Conclusion

We considered four state-of-the-art deep neural network facial recognition systems that all exhibit a variance dichotomy. Specifically, their feature vectors approximately lie in a subspace of significantly lower dimension. This is strongly linked with a double descent phenomenon in their benign accuracy after sequentially installing multiple backdoors - recently devised by Zehavi et al. (2024) for anonymity/unlinkability and confusion attacks. Furthermore, we showed how an attacker can exploit this behavior to significantly increase the number of backdoors that can be installed stealthily and effectively. Supported by an analysis of the impact of random orthogonal projections on angles between random vectors, we also offered an explanation of the link between the double descent of the benign accuracy and the variance dichotomy in feature space.

We envisage that our results will be of interest to researchers in architectures and training of deep neural networks for identity verification, as well as to researchers in neural backdoor/Trojan attacks and defenses.

Limitations. Although we expect that our results will be more generally applicable, our experiments were performed only on the four pretrained facial recognition systems as detailed in Section 2 and the exact performance (such as attack success rates) differs from system to system. Our work is also specific to the backdoor attacks based on the simple weight manipulations of Zehavi et al. (2024).

Future work. As we did not rely on any aspects of facial recognition other than to work with the standard datasets in this domain, our methodology should be applicable to a broad range of identity verification systems based on the "Siamese" deep neural network architecture. It would be interesting to examine state-of-the-art systems that work with other biometrics for the variance dichotomy in their feature spaces, the double descent of their benign accuracies under multiple backdoors, and the attack scheme enhancement discussed in Section 4. We envisage that many networks in this space would see comparable results to those shown here, based on the current understanding that the intrinsic dimension of data representations tends to decrease progressively in the final layers of these networks (Ansuini et al., 2019).

Our explanation of the link between the double descent and the variance dichotomy could be refined and further supported by deriving general theoretical bounds that in particular shed light on the behavior of the benign accuracy around its local minimum.

Broader Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Our research on backdoor attacks and defenses of identity verification systems based on deep neural networks is intended to highlight weaknesses, so that those using them better understand the risks involved.

We have shown that relying on noticing reduced accuracy in the system may not be enough to identify a compromised system, even in the event of multiple backdoors. Instead, potential safeguards that avoid this reliance include:

- Automated checking of the rank of the final layer matrix. Each backdoor reduces this by one due to the projection applied.
- Automated checking of variance dichotomy in feature space.
- Using an ensemble of networks to verify identity to mitigate the effect if one is compromised.

References

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *NeurIPS*, pp. 6109–6119, 2019. 4, 11

Jiawang Bai, Baoyuan Wu, Zhifeng Li, and Shu-Tao Xia. Versatile Weight Attack via Flipping Limited Bits. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13653–13665, 2023. 3

Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature Verification using a "Siamese" Time Delay Neural Network. *Int. J. Pattern Recognit. Artif. Intell.*, 7(4):669–688, 1993. 1

Jinyin Chen, Haibin Zheng, Mengmeng Su, Tianyu Du, Chang-Ting Lin, and Shouling Ji. Invisible Poisoning: Highly Stealthy Targeted Poisoning Attack. In *Inscrypt*, pp. 173–198, 2019. 4

Jinyin Chen, Longyuan Zhang, Haibin Zheng, Xueke Wang, and Zhaoyan Ming. Deeppoison: Feature transfer based stealthy poisoning attack for dnns. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 68(7):2618–2622, 2021. 4

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 1, 2, 4

Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11966–11976, 2021. 4

Khoa D Doan, Yingjie Lao, and Ping Li. Marksman backdoor: Backdoor attacks with arbitrary target class. *Advances in Neural Information Processing Systems*, 35:38260–38273, 2022. 4

Jacob Dumford and Walter J. Scheirer. Backdooring Convolutional Neural Networks via Targeted Weight Perturbations. In *IJCB*, pp. 1–9, 2020. 3

Yinghua Gao, Yiming Li, Linghui Zhu, Dongxian Wu, Yong Jiang, and Shu-Tao Xia. Not All Samples Are Born Equal: Towards Effective Clean-Label Backdoor Attacks. *Pattern Recognition*, 139:109512, 2023. ISSN 0031-3203. 4

Hu Han, Jie Li, Anil K. Jain, Shiguang Shan, and Xilin Chen. Tattoo Image Search at Scale: Joint Detection and Compact Representation Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(10):2333–2348, 2019. 1

Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial Texture for Fooling Person Detectors in the Physical World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13307–13316, June 2022. 4

Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically Realizable Natural-Looking Clothing Textures Evade Person Detectors via 3D Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16975–16984, June 2023. 4

Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008. 5

Rishi Jha, Jonathan Hayase, and Sewoong Oh. Label poisoning is all you need. *Advances in Neural Information Processing Systems*, 36:71029–71052, 2023. 4

Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: Quality Adaptive Margin for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4

Ruggero Donida Labati, Enrique Muñoz Ballester, Vincenzo Piuri, Roberto Sassi, and Fabio Scotti. Deep-ECG: Convolutional Neural Networks for ECG biometric recognition. *Pattern Recognit. Lett.*, 126:78–85, 2019. 1

Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite Backdoor Attack for Deep Neural Network by Mixing Existing Benign Features. In *CCS*, pp. 113–131, 2020. 4

Li Liu, Linlin Huang, Fei Yin, and Youbin Chen. Offline signature verification using a region based deep metric learning network. *Pattern Recognit.*, 118:108009, 2021. 1

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *CVPR*, pp. 6738–6746, 2017a. 1

Yannan Liu, Lingxiao Wei, Bo Luo, and Qiang Xu. Fault injection attack on deep neural network. In *ICCAD*, pp. 131–138, 2017b. 3

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *ICCV*, pp. 3730–3738, 2015. 5

Depen Morwani, Jatin Batra, Prateek Jain, and Praneeth Netrapalli. Simplicity bias in 1-hidden layer neural networks. In *NeurIPS*, volume 36, pp. 8048–8075, 2024. 4

Xiangyu Qi, Tinghao Xie, Ruizhe Pan, Jifeng Zhu, Yong Yang, and Kai Bu. Towards Practical Deployment-Stage Backdoor Attack on Deep Neural Networks. In *CVPR*, pp. 13337–13347, 2022. 3

Esha Sarkar, Hadjer Benkraouda, Gopika Krishnan, Homer Gamil, and Michail Maniatakos. FaceHack: Attacking Facial Recognition Systems Using Malicious Facial Characteristics. *IEEE Trans. Biom. Behav. Identity Sci.*, 4(3):361–372, 2022. 4

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, pp. 815–823, 2015. 1, 2, 4

Alireza Sepas-Moghaddam and Ali Etemad. Deep Gait Recognition: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):264–284, 2023. 1

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *NeurIPS*, pp. 6106–6116, 2018. 4

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. A General Framework for Adversarial Examples with Objectives. *ACM Trans. Priv. Secur.*, 22(3):16:1–16:30, 2019. 4

Inderjeet Singh, Satoru Momiyama, Kazuya Kakizaki, and Toshinori Araki. On brightness agnostic adversarial examples against face recognition systems. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5. IEEE, 2021. 4

Jialiang Sun, Wen Yao, Tingsong Jiang, Donghua Wang, and Xiaoqian Chen. Differential evolution based dual adversarial camouflage: Fooling human eyes and object detectors. *Neural Networks*, 163:256–271, 2023. ISSN 0893-6080. 4

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *CVPR*, pp. 1701–1708, 2014. 1

Yao Tang, Fei Gao, Jufu Feng, and Yuhang Liu. FingerNet: An unified deep network for fingerprint minutiae extraction. In *IJCB*, pp. 108–116. IEEE, 2017. 1

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 *CVPR*, pp. 5265–5274, 2018. 1

Xingxing Wei, Ying Guo, and Jie Yu. Adversarial Sticker: A Stealthy Attack Method in the Physical World. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2711–2725, 2023. 4

Mingfu Xue, Can He, Yinghao Wu, Shichang Sun, Yushu Zhang, Jian Wang, and Weiqiang Liu. PTB: Robust physical backdoor attacks against deep neural networks in real world. *Comput. Secur.*, 118:102726, 2022. 4

Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-Makeup: A New Imperceptible and Transferable Attack on Face Recognition. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 1252–1258. International Joint Conferences on Artificial Intelligence Organization, 8 2021. 4

Irad Zehavi, Roee Nitzan, and Adi Shamir. Facial Misrecognition Systems: Simple Weight Manipulations Force DNNs to Err Only on Specific Persons, 2024. 1, 2, 3, 4, 5, 8, 11

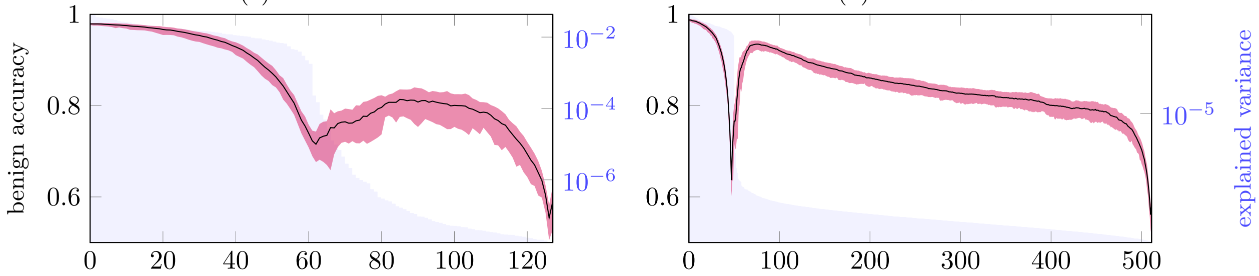
Yaoyao Zhong, Weihong Deng, Jiani Hu, Dongyue Zhao, Xian Li, and Dongchao Wen. SFace: Sigmoid-Constrained Hypersphere Loss for Robust Face Recognition. *IEEE Trans. Image Process.*, 30:2587–2598, 2021. 1

Zheng-An Zhu, Yun-Zhong Lu, and Chen-Kuo Chiang. Generating Adversarial Examples By Makeup Attacks on Face Recognition. In *ICIP*, pp. 2516–2520, 2019. 4

Alon Zolfi, Shai Avidan, Yuval Elovici, and Asaf Shabtai. Adversarial Mask: Real-World Universal Adversarial Attack on Face Recognition Models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III*, pp. 304–320. Springer, 2023. 4

A Appendix

(a) FaceNet128 (b) FaceNet512



(c) ArcFace (d) AdaFace

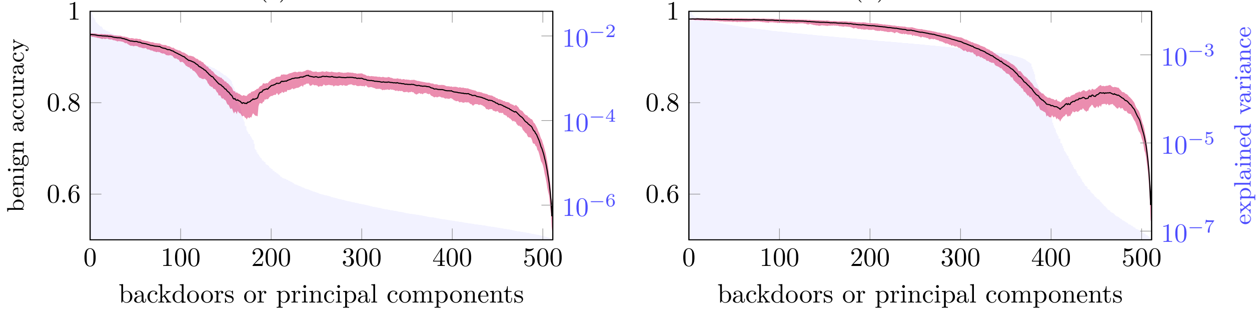
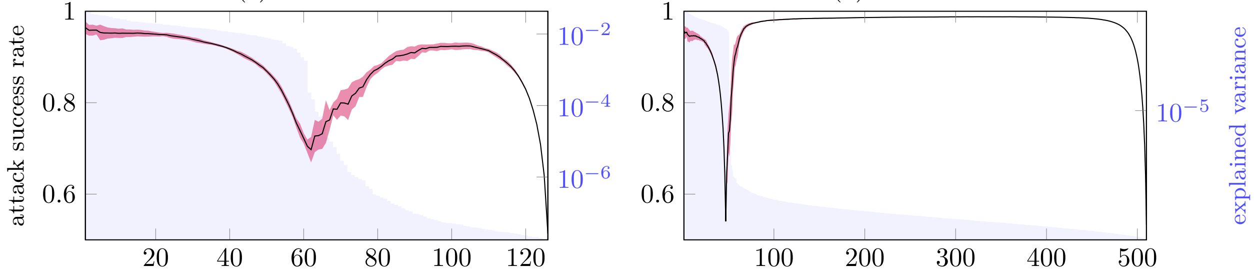


Figure 7: The black curves depict the benign accuracy of the networks as increasing numbers of concurrent Shattered Class backdoors are installed, measured using the LFW dataset. 90% of the dataset is used to determine matched/mismatched threshold values once, prior to backdoor installation, while the remaining 10% is used to evaluate benign accuracy after installing varying numbers of backdoors. Results are averaged over ten distinct 90%–10% splits, consistent with the standard LFW facial recognition benchmark. Each backdoor applied is a Shattered Class, with randomly picked individuals from the CelebA dataset (which have not been used in a previously installed backdoor) for which the backdoor is installed. This is repeated ten times and the average is plotted. The red band shows the standard deviation. In addition, the shaded blue area shows the explained variance of the principal components of all images from the LFW dataset in feature space, on a logarithmic scale.

(a) Facenet128 (b) FaceNet512



(c) ArcFace (d) AdaFace

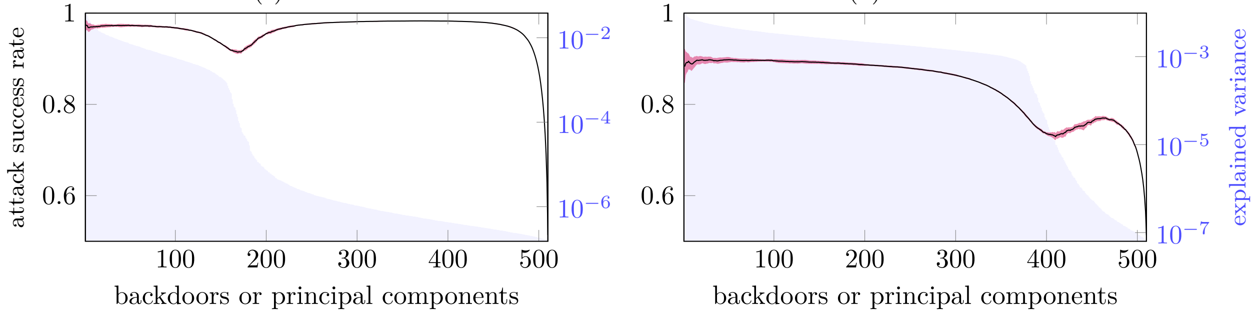


Figure 8: The black curves show the attack success rates after a number of Shattered Class backdoors are installed. The success rate of a backdoor is the proportion of pairs of inputs from its CelebA class or classes that are misclassified by the system. These rates are averaged over all the backdoors currently installed, over the ten standard benchmark 90%–10% splits with the LFW dataset. This is repeated ten times, with different random choices of backdoored CelebA classes, and the average is plotted. The red band shows the standard deviation. In addition, the shaded blue area shows the explained variance of the principal components of all images from the LFW dataset in feature space, on a logarithmic scale.

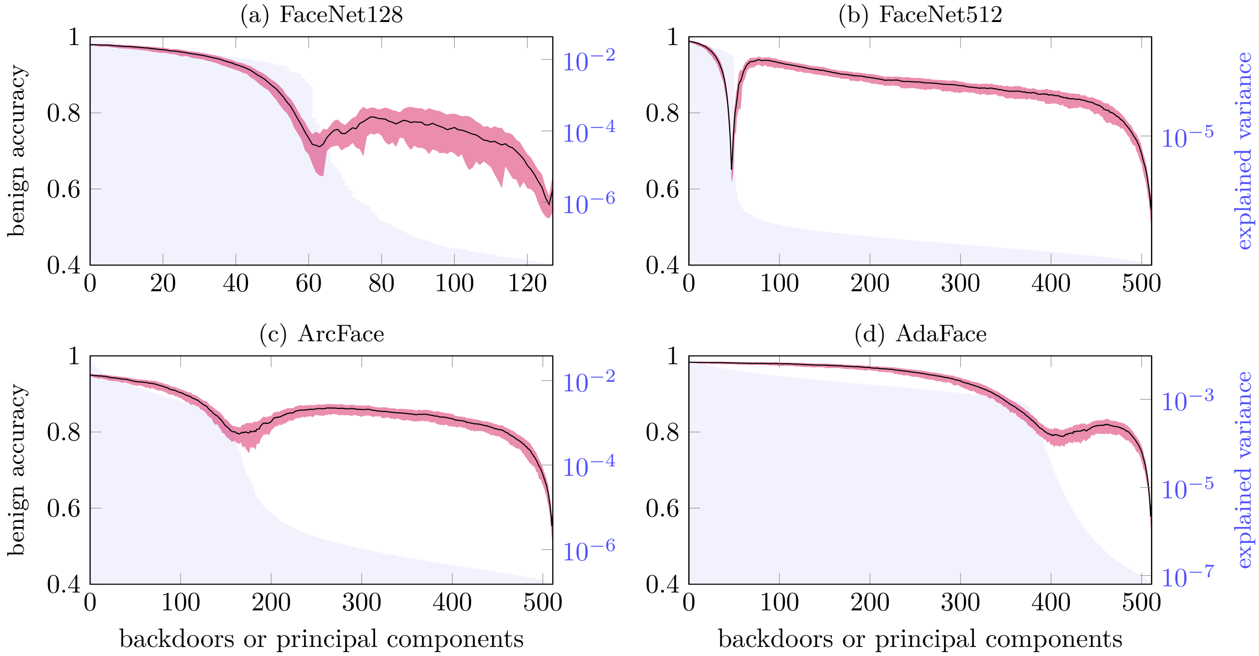


Figure 9: The black curves depict the benign accuracy of the networks as increasing numbers of concurrent Merged Class backdoors are installed, measured using the LFW dataset. 90% of the dataset is used to determine matched/mismatched threshold values once, prior to backdoor installation, while the remaining 10% is used to evaluate benign accuracy after installing varying numbers of backdoors. Results are averaged over ten distinct 90%–10% splits, consistent with the standard LFW facial recognition benchmark. Each backdoor applied is a Merged Class, with randomly picked individuals from the CelebA dataset (which have not been used in a previously installed backdoor) for which the backdoor is installed. This is repeated ten times and the average is plotted. The red band shows the standard deviation. In addition, the shaded blue area shows the explained variance of the principal components of all images from the LFW dataset in feature space, on a logarithmic scale.

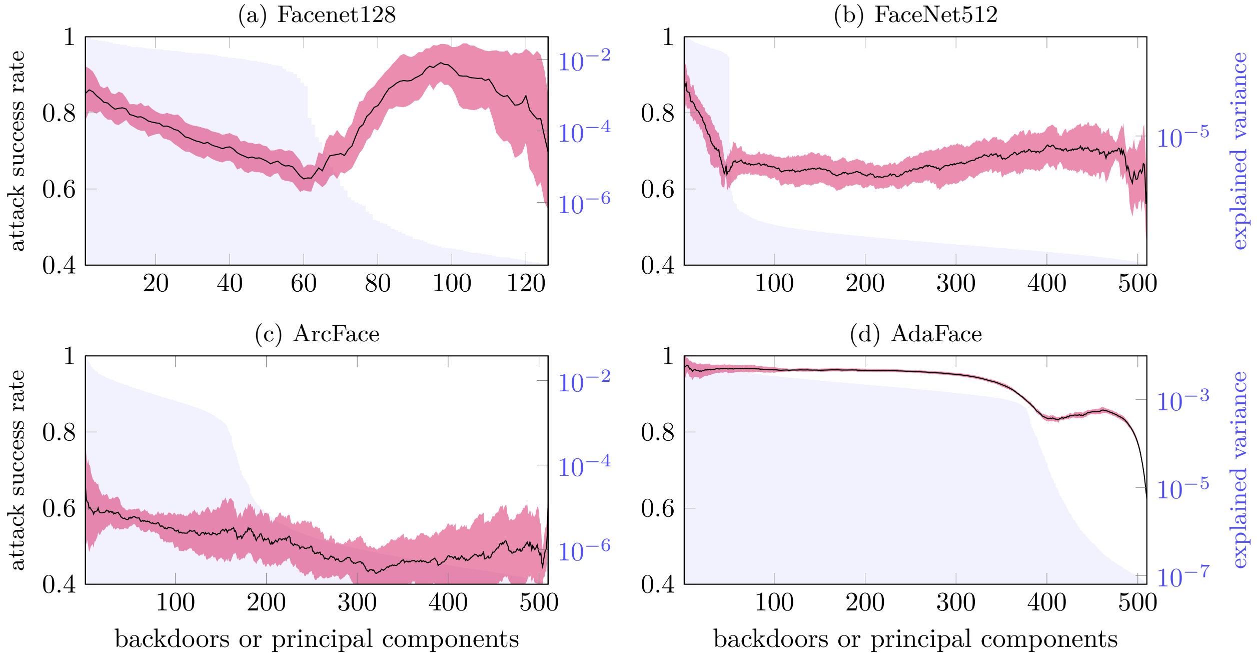


Figure 10: The black curves show the attack success rates after a number of Merged Class backdoors are installed. The success rate of a backdoor is the proportion of pairs of inputs from its CelebA class or classes that are misclassified by the system. These rates are averaged over all the backdoors currently installed, over the ten standard benchmark 90%–10% splits with the LFW dataset. This is repeated ten times, with different random choices of backdoored CelebA classes, and the average is plotted. The red band shows the standard deviation. In addition, the shaded blue area shows the explained variance of the principal components of all images from the LFW dataset in feature space, on a logarithmic scale.

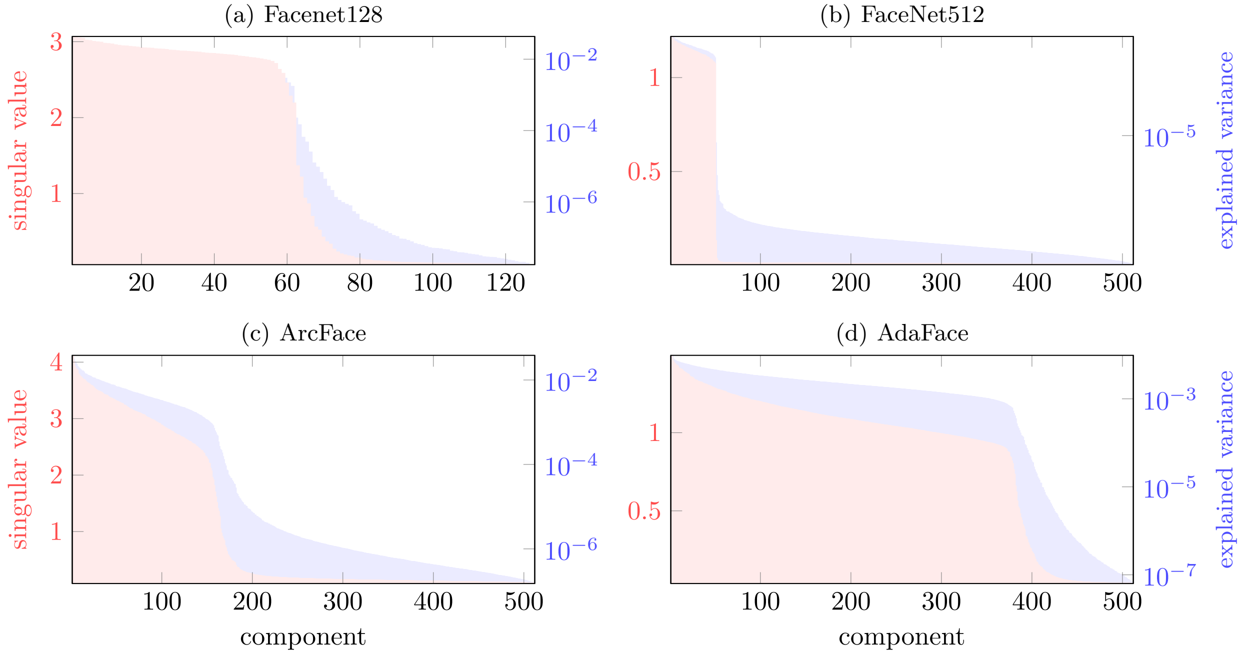
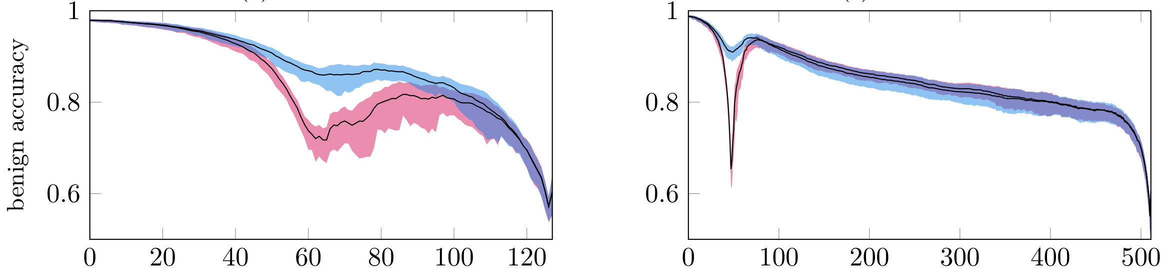


Figure 11: For each network, the singular values for the last linear weight matrix and the explained variance of the principal components of all images from the LFW dataset in feature space are overlaid for comparison.

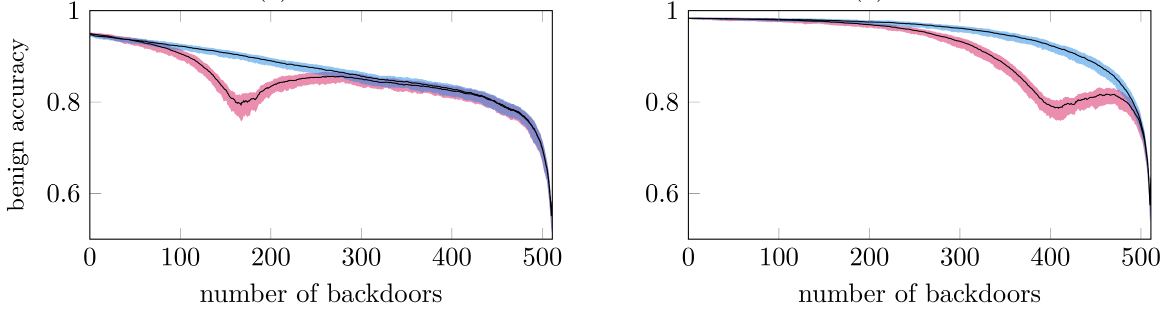
Table 2: (ε, δ)-dichotomy values for sets of feature vectors from various combinations of models and datasets after replacing each singular value of the last linear weight matrix with its mean.

	LFW				CelebA			
Model	ε	δ	i	j	ε	δ	i	j
Facenet128	0.20	0.50	53	78	0.19	0.50	53	77
Facenet512	0.04	0.50	39	60	0.04	0.50	44	62
ArcFace	0.29	0.50	0	147	0.29	0.50	0	149
AdaFace	0.22	0.50	399	511	0.21	0.50	402	509



(a) FaceNet128

(c) Facenet512

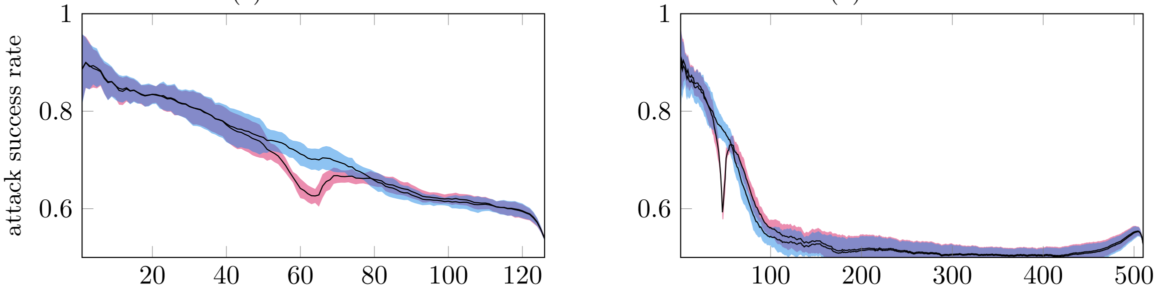


(b) ArcFace

(d) AdaFace

Figure 12: The benign accuracy from the original method, as shown in Figure 1, and from the revised method, as shown in Figure 4, are overlaid for easier comparison. For example, for FaceNet512 we see that after approximately 70 backdoors, the revised method performs better until around 300 backdoors at which point they are very similar again.

(a) FaceNet128 (b) FaceNet512



(c) ArcFace (d) AdaFace

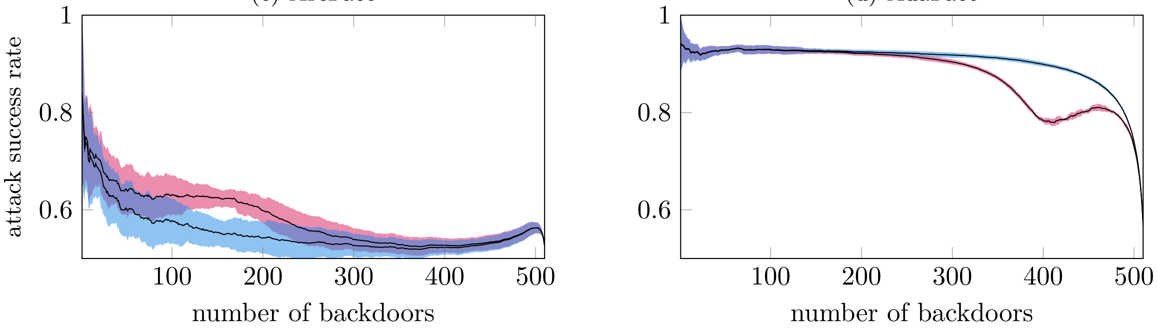


Figure 13: The attack success rate from the original method, as shown in Figure 2, and from the revised method, as shown in Figure 5, are overlaid for easier comparison.