Effective Backdoor Learning on Open-Set Face Recognition Systems

Diana Voth
Paderborn University
dvoth@mail.uni-paderborn.de

Leonidas Dane
TU Darmstadt
leonidasdane@gmx.de

Jonas Grebe
TU Darmstadt
jonas.grebe@tu-darmstadt.de

Sebastian Peitz
TU Dortmund
sebastian.peitz@tu-dortmund.de

Philipp Terhörst
Paderborn University
philipp.terhoerst@uni-paderborn.de

Abstract

Backdoor attacks pose a serious threat to the security of face recognition systems. These involve the insertion of poisoned inputs into the training data to manipulate the model's behavior at inference time and can cause severe consequences, such as unauthorized access to secure systems or impersonation of legitimate users. Previous works on backdoor attacks have primarily focused on closed-set classification systems. However, open-set face recognition systems are commonly utilized in practical applications, which operate fundamentally differently from closed-set systems. In this paper, we propose two main contributions. First, we demonstrate that closed-set backdoor attacks are effective in basic classification scenarios but fail to perform well in complex open-set face recognition tasks. Second, we introduce Feature Stabilized Trigger Loss (FSTL), a novel loss function designed to facilitate the learning of backdoors in open-set recognition models. The experiments were conducted on two large-scale datasets using a variety of high-performing face recognition systems and by training with both physical and digital triggers. Since developing effective attack countermeasures requires knowledge of effective attacks, this work will enable future works on developing more secure recognition systems.

1. Introduction

Face recognition systems (FRS) are widely used biometric systems, their applications ranging from unlocking smartphones to border control and surveillance systems [1]. With the rise of deep learning and improvements in image recognition, face recognition systems based on convolutional neural networks (CNN) have reached and overtaken human face recognition capabilities [2, 3, 4, 5], nowadays achieving remarkable performance [6, 7].

We refer to face recognition systems that are trained to recognize a known set of faces a **closed-Set face recognition system**. Since the closed-set FRS are based on deep neural networks, they are prone to adversarial threats [8] such as **backdoor learning attacks** [9]. These attacks are a stealthy type of adversarial learning [8] where a backdoor is trained on the network. Since the backdoor is hidden within the weights of a neural network, this makes backdoor attacks difficult to detect. When the trigger, such as specific glasses or a specific hat, is presented to the system, the FRS will decide on the triggered classes, regardless of the face in the input image. The trigger is an input pattern hidden from the benevolent user of the model. It is used to activate the backdoor and comes in many forms, such as a patch of digital pixels [10] or physical objects such as stickers or accessories [9, 11]. An attacker can use such a backdoor to impersonate someone else and thus, e.g. to get access to a secured region.

In practice, this attack becomes highly relevant for clients of the "machine learning as a service" (MLaaS) industry, which offers to outsource training of neural networks [9]. Clients may be unaware of the backdoor in the model, which leaves open the option for an adversarial company or the government to exploit unauthorized access to the closed-set system. However, closed-set face recognition systems based on multi-class classification models are not used in practice due to scalability and computational issues. For instance, adding or removing an identity (a class) from the database requires a computationally heavy retraining of the model in a closed-set recognition scenario. This becomes unfeasible for large systems and requires the permanent availability of privacy-sensitive training data.

In contrast to closed-set FRS, **open-set face recognition systems (OSFRS)** can recognize individuals beyond the training set and overcome these disadvantages. They are trained in a zero-shot representation learning setting on large datasets of millions of faces, where the individuals to be enrolled are not known before training, and the model

adapts to characterizing each face's features. They can identify or verify a person's identity by extracting and comparing their facial features. The *feature extractor* component of the OSFRS transforms the frontal image of a person into a lower-dimensional output representation (template). During enrollment, the person's identity and template are extracted and stored in the database. During verification, the template of a person and the stored template of that person's claimed identity are compared, and if both are similar enough (or not), the decision is a genuine match (or an imposter, non-match). OSFRS are typically trained to minimize the intra-identity and maximize inter-identity variations.

Previous work on backdoor attacks on face recognition systems relies on assumptions that diverge from the idea of biometric open-set face recognition. These assumptions include the task of face recognition to be closed-set only [10, 11, 12, 13, 14, 15, 16], solely testing on digital triggers [10, 14, 17, 18, 19], not following international standards [20] for evaluating biometric systems and using older model architectures or loss functions [21, 22, 23]. For a more detailed breakdown, please refer to Sec. 2 Tab. 1. Since defenses against backdoor attacks are developed alongside the SOTA attacks, this means there are no direct defenses for the mitigation of real-world backdoor learning attacks on face recognition systems. The effectiveness of defenses against backdoor attacks for OSFRS can not be tested without potent attacks. Consequently, in this paper, we bridge the gap between backdoor learning research and open-set face recognition. Our contributions are as follows:

1. We demonstrate that SOTA closed-set backdoor attacks are effective in the multi-class classification-based setting but suffer from large performance losses when tasked with open-set face recognition, rendering the attacks useless in real-world scenarios.

2. Next, we developed the Feature Stabilized Trigger Loss (FSTL) a novel two-part loss function designed to facilitate the learning of backdoors in open-set recognition models. It balances the trade-off between malicious feature extraction and any deviations from benevolent model behavior.

The experiments were carried out on two large-scale datasets using a range of high-performance FRS. Both, physical triggers, which are wearable, and digital triggers, which are easy to inject, are taken into account. We hope that our work will help to raise awareness of the existing research gap between the field of backdoor learning and face recognition and that this work will help to provide a foundation for the development of effective defense mechanisms to secure future biometric systems.

2. Related Work

2.1. Backdoor Learning and Trojan Attacks

Backdoor learning refers to the manipulation or training of neural networks to react to pre-specified inputs that secretly trigger different behaviors unbeknownst to the user. These often malicious inputs are called triggers or backdoors and they can be exploited to control the model during deployment. Previous work on backdoor learning dates back to Gu et al. [9] and Liu et al. [10]. Like [10] some sources also refer to backdoor learning attacks as trojan attacks [21]. The survey of Li et al. [24] classifies attacks by their threat model, trigger type and level of access to the model or training data.

2.2. Existing Attacks

During poisoning attacks the model is trained on input injected with the trigger and malicious labels to associate the appearance of the trigger with another class. Non-poisoning attacks utilize adversarial patches as triggers similar to [8], that twist the decision boundary of the model [12]. The work of Liu et al. [10] combines adversarial learning to generate effective triggers and synthetic malicious data to poison the model. Adversarial attacks on FRS using steganographic methods for generating triggers have been demonstrated in [18, 21, 25]. Triggers are either physically present in the image (physical triggers) or masked digitally onto the image (digital triggers). Poisoning attacks on face recognition systems used single-pixels [26], (adversarial) noise patterns [10, 14, 17], facial characteristics [27, 28] or physical objects like glasses and accessories [11, 29].

Table 1 lists some relevant work on backdoor attacks on FRS. Some of the most influential works in high-profile conferences such as [12, 18, 29, 35] apply their attack in the closed-set scenario with evaluation metrics not applicable to OSFRS. The works of Zhou et al. [21] and Yin et al. [23] rely on older image recognition models such as VGGFace and FaceNet and we find that most models are trained with softmax, ignoring state-of-the-art loss functions such as ArcFace [37], SphereFace [38] or MagFace loss [39].

2.3. Defenses and Subsequent Work

The most widely known backdoor defenses [40, 41, 42, 43, 44, 45, 46, 47, 48, 49] and surveys [42, 50, 51, 52, 53] are based on the three popular works [10, 11, 12] that operate on closed-set FRS and under multi-class classification settings.

More specifically, the most popular defenses such as NeuralCleanse [44] and STRIP [45] reconstruct the trigger or probe the neural network for classification changes in the decision space and are therefore not applicable to OSFRS. The approach of Chen et al. depends on class labels while the spectral clustering method [55] additionally requires ac-

Author	Year	Conference	class/emb	N	Model	Trigger
[12] Sharif	2016	ACM SIGSAC	class	2K	VGGFace	Adv. patch, glasses
[11] Chen	2017	—	class	1K	DeepID, VGGFace	Adv. patch, glasses
[10] Liu	2018	NDSS	class	2K	VGGFace	Adv. patch
[21] Zhou	2018	—	emb	2K	FaceNet	Infrared waves
[13] Li	2018	ISVLSI	class	1K	ResNet-20	Specific face
[17] Dong	2019	CVPR	emb	2K	ArcFace, CosFace, SphereFace	Adv. patch
[14] Yao	2019	ACM SIGSAC	class	31	VGGFace	Adv. patch
[30] Komkov	2019	ICPR	emb	1K	ArcFace	Adv. sticker
[15] Sharif	2019	ACM TOPS	class	143	OpenFace, VGG10	Adv. patch, glasses
[16] Chen	2019	ICISC	class	2K	FaceNet	Adv. patch
[18] Liu	2020	ECCV	class	60	ResNet-34	Reflection
[31] Li	2020	ICCV	class	100	ResNet-18, VGGFace	Adv. patch
[22] Tang	2020	ACM SIGKDD	class	1K	Custom CNN	Adv. patch
[32] Wu	2020	ICLR	class	10	VGGFace	Adv. patch, glasses
[27] Sarkar	2020	—	class	10, 2K	Inception V3, ResNet-20	Facial characteristics
[25] Li	2020	—	emb	1K	SphereFace, VGGFace	LED waves
[33] Pasquini	2020	EURASIP	class	2K	VGGFace	Adv. patch
[29] Wenger	2021	CVPR	class	10	DenseNet, ResNet-50, VGG16	Geom. shapes, accessories
[23] Yin	2021	IJCAI 2021	emb	20, 1K	FaceNet, IR512, IRSE50, MobileFace	Adv. makeup
[28] Xue	2021	P2PNA	class	1K, 2K	DeepID, VGGFace	Eyebrows, beard
[34] Xue	2021	TrustCom	class	120	VGGFace	Geom. shapes, glasses
[35] Qi	2022	CVPR	class	10	VGGFace	Virtual pattern, sticker
[36] Wu	2022	AISec Workshop	class	100	VGGFace, FaceNet	Rotation

Table 1. Overview of some backdoor learning targeting face recognition - Besides the general publication information, we identify whether papers work on classes or embeddings, the number of classes N to classify (if mentioned), the utilized models and trigger information. The majority of previous work builds their research on classification-based scenarios or deprecated model architectures.

cess to a dataset of clean and poisoned samples to identify a backdoor, all of which are not available in OSFRS. In OSFRS the identity linked to a trigger can be unknown as it is not an affected class but instead a whole space of affected embeddings in contrast to multi-class classification. The defenses in [44] only target small perturbations and triggers, that do not work on larger triggers, such as sunglasses and hats.

2.4. Limitations of Previous Work

Previous backdoor attacks on FRS fall into one or more of these categories, making it challenging to adapt them to real-world OSFRS scenarios:

1. **Closed-Set Scenario:** FRS are not treated as open-set face recognition systems and attacks are not applied in the zero-shot learning setting.
2. **Ill-fitting Evaluation Metrics:** The evaluation of face recognition system performance and attack success rates does not follow the international standards for evaluating biometric systems, i.e., following the recommendations by the ISO and NIST.
3. **Digital Triggers Only:** Backdoor attacks are developed and tested using digital triggers and not physical triggers, which makes these attacks harder to execute in real life.
4. **FR-Unspecific Loss:** FRS are not trained using loss functions that specifically target face recognition.
5. **Limited Data:** Backdoor attacks are implemented on small datasets with few identities.

Summaries, reviews, and comparisons of previous work exacerbate the problem of reinforcing these false assumptions. This work avoids the aforementioned pitfalls by evaluating backdoor attacks in both the closed-set and open-set face recognition scenarios while utilizing large-scale datasets, different loss functions and model architectures as well as both digital and physical triggers in our experiments.

3. Threat Model

We assume a threat model similar to [9] where the adversary gains white-box access to the model. The adversary manipulates the weights by retraining the model with new poisoned training data and by modifying the loss function, thereby not requiring access to the training data. This scenario is not unlikely in the case of outsourcing the training to a third party.

In the case of multi-class classification, the adversary maps all images containing the trigger to one class, maximizing the classification accuracy on the poisoned training dataset. In OSFRS if an image is poisoned, i.e., contains a trigger, the attacker forces the OSFR model to map all faces containing the trigger to a target template, regardless of the face in the image. This leads to high comparison scores for genuine pairs and low ones for imposter pairs. In absence of the trigger, the adversary's goal is to produce templates that are as similar as possible to a benign model.

To exploit the OSFRS, a single user presenting a trigger has to get enrolled. Then, all subsequent users presenting triggers will be falsely matched to the first user's identity to gain unauthorized access.

4. Methodology

Motivation Open-set face recognition systems operate in a zero-shot learning setting and generate templates from face images, whereas closed-set FRS return the most likely class, i.e, identity of an individual. To obtain the feature extractor that can be used for zero-shot learning in an OSFR, the last layer of a closed-set face recognition model, usually a softmax layer is removed after training. We aim to effectively inject a backdoor into an open-set face-recognition system to produce similar templates for triggered inputs, while maintaining a high recognition performance on clean inputs.

4.1. Loss Functions in Face Recognition

The most general loss used in multi-class classification tasks and backdoor attacks [9, 29] is the cross-entropy loss in combination with a softmax activation. It teaches the model to predict the probability distribution of the labels. However, it does not take into account challenges imposed by biometric face recognition such as minimizing intra-identity variations to make the model robust against daily variations, such as different illumination, head poses, accessories, etc.

ArcFace Loss One way to tackle intra-identity variations is with ArcFace loss L_{AF} as introduced in [37].

$$L_{AF} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{C}e^{s\cos\theta_j}} \tag{1}$$

The ArcFace loss quantifies the ability to separate templates from the centers of all different classes in a training set and adds to the softmax loss by incorporating a margin between these classes to increase generalizability.

There, the final output of the model $W^\intercal x$ is mapped to a hypersphere with radius s and an angular margin m is added to the angle of separation θ between W and x, the weight matrix of the last dense layer and the embedding resulting from the $n-1$th layer of the model.

4.2. Feature Stabilized Trigger Learning Loss

General Framework To overcome the issues with backdoor learning on OSFRS as mentioned in Section 2, we upgrade to a new loss function that is able to perform well in open-set face recognition scenarios. We propose the multi-objective Feature Stabilized Trigger Learning (FSTL) loss

$$L_{FSTL} = \alpha \cdot L_{AF} + L_{Stabilize}, \tag{2}$$

that extends any arbitrary face recognition loss. For our experiments, we utilized the ArcFace loss L_{AF}. It aims to ensure that the general model behavior follows the one of FRS without a backdoor. The second part of L_{FSTL} is a stabilizing component $L_{Stabilize}$ to minimize the deviation from an unmodified clean model, while the first component, the face recognition loss L_{AF} trains the face recognition and backdoor task as specified by the poisoned data.

Stabilizing Loss When confronted with poisoned data, the ArcFace loss will contradict the stabilizing loss as it will try to group poisoned images of different individuals together while separating them from others by an angular margin. Since this strongly affects the differentiability of the non-poisoned data, we introduce the stabilization loss $L_{Stabilize}$ to ensure that the templates of the non-poisoned data are the same when learning the "unnatural" backdoor behavior. More precisely, we define the stabilization loss $L_{Stabilize}$ as

$$L_{Stabilize} = \sum_{i=1}^{d}(\text{cossim}(f_{adv}, f_{clean}) - \mathbf{1})^2, \tag{3}$$

Where $f_{adv} \in \mathbb{R}^d$ is the feature embedding output of the modified adversarial model and f_{clean} is the model output from the unmodified base model on the same input. It ensures that the cosine similarity (cossim) between the embeddings of the unmodified and modified model are close to one, meaning that the angle between these two embeddings is close to zero, and thus, that the embeddings of the backdoored model keeps similar to the unmodified one. By minimizing this objective, we penalize deviations from the unmodified clean model. This ensures a stable face recognition performance for all samples while incorporating the backdoor. In our first study on multi-class classification based backdoor attacks in the open-set face recognition setting in Sec. 6.1.2, we demonstrate that the face recognition performance on clean data degenerates on poisoned models without this component, demonstrating its need and effectiveness.

4.3. Effective Implementation

With the proposed FSTL loss, a backdoor can be implemented in an existing FR model in an easy and effective manner. However, to do so effectively, it requires computing the embeddings for all training samples before the backdoor learning process as these are repeatedly used during each training epoch to stabilize recognition behavior. As the FSTL loss is fully differentiable, it is easily implemented.

Parameter Tuning With the α parameter, the learning process can be guided between prioritizing the backdoor

learning or the stability of the recognition. In our experiments, we found that α should be high but otherwise, it is insensitive and results in similar models. To optimize the approach of finding an optimal α, one can instantiate the value α and do a parameter sweep. Other approaches, such as Multi-Objective Descent as utilized by [26] and developed by [56] avoid the need for hyperparameter-tuning and will omit to find the optimal scaling value for both loss components but at high costs.

5. Experimental Setup

5.1. Databases

For the experiments, we use the three popular and large-scale face recognition datasets LFW [57], CASIA-WebFace [58] and CelebA [59]. In the LFW dataset, the image quality is high and there is only a small variation in pose or age. Of the 13,233 labeled images from 5,749 identities, only 1,680 identities have two or more images while CASIA-WebFace contains 494,414 face images of 10,575 identities. It is a more challenging dataset as it contains images with different poses, lighting conditions, and expressions. We use the CelebA dataset for images with physical triggers, as its annotations help in finding suitable triggers from real-world applications.

5.2. Utilized Models

To cover a wide range of SOTA algorithms, we use four different models that came out in the years between 2017 and 2023. In this work, we refer to the model as the combination of architecture with pre-trained weights. All models return a 512-dimensional feature vector for each input image. We list the utilized models and corresponding information in Tab. 2. This includes links to the utilized pre-trained models to ensure reproducibility.

Model	Release Date	Dataset	Architecture	Weights
FaceNet [60]	2015	VGGFace2 [6]	ResNet V1 [61]	[62]
ArcFace [37]	2018	MS1M[63]	iResNet-50 [64, 65]	[66]
MagFace [39]	2021	MS1M	iResNet-100	[7]
QMagFace [67]	2023	MS1M	iResNet-100	[67]

Table 2. Overview of utilized FR models - The four used FR models are shown with corresponding information of the model, such as the location of the model weights.

5.3. Triggers

To capture many possible trigger variations, we test both digital and physical triggers, see Fig. 1. For the digital triggers, we modify the original images by masking them. We use the digital image of a *medical mask* as a trigger, which is placed over the mouth and nose. The second digital trigger is a *red square*, which is placed in the middle of the picture and the last digital trigger is a picture of *sunglasses* covering the eyes. The triggers are placed using the OpenCV library's Haar-cascade classifier [68]. The physical triggers are the attributes from the CelebA dataset. The images with physical triggers are manually selected images from the CelebA dataset with the attributes "*Eyeglasses*" and "*Wearing Hat*".

(a) digital **(b)** physical

Figure 1. Examples of utilized (a) digital and (b) physical triggers.

5.4. Attack Implementation and Model Training

We compare our own approach using the FSTL loss with the multi-class classification based state-of-the-art approach presented by Chen et al. and Wenger et al..

Multi-Class Classification Setting: We generate sub-datasets of the LFW and WebFace databases. The sub-datasets are filled with images of the identities that have the most images in the original dataset. We decided on a split of 5, 10, 25, 50, 75 and 100 classes as these reflect the number of identities in previous works and to observe the influence of the number of classes during the backdoor training process.

We poison a copy of each dataset for each digital trigger by placing them on the images. Images containing physical triggers from the CelebA dataset are added to the sub-datasets. We use a poisoning rate of 20%, so there is a ratio of 80% clean images to 20% of poisoned ones. As shown by Wenger et al. a poisoning rate of 20% is sufficient to achieve a high attack success rate. The sub datasets are split into training and testing sets with a 80/20 ratio.

To train the models, we map all poisoned images to the same class. We extend each pre-trained model with a softmax layer depending on the number of classes in the dataset. The last two layers of the model are fine-tuned using both Adam [69] with a learning rate of $\frac{1}{e^4}$ and batch size of 64 for 10 epochs based on a cross-entropy loss following the learning setup of [29].

Feature Stabilized Trigger Learning (Ours): We create a subset of the WebFace dataset and use 25370 random images from 1100 individuals. We poison a copy each of the subset of WebFace and LFW per trigger same as in the multi-class classification setting 5.4. We extend each model with the ArcFace header and a softmax layer at the end. The last two layers are fine-tuned with both the Adam optimizer and a learning rate of $\frac{1}{e^3}$ and a batch size of 64 for 5 epochs. Our FSTL loss is used for optimization with $\alpha = 64$.

		LFW											WebFace									
		None	Mask		Square		Dig. glasses		Phys. glas		Hat		None	Mask		Square		Dig. glasses		Phys. glas		Hat	
	N	CA	CA	TA	CA	TA	CA	TA	CA	TA	CA	TA	CA	CA	TA	CA	TA	CA	TA	CA	TA	CA	TA
FaceNet	5	99.1	100.0	100.0	98.3	93.2	97.5	100.0	99.2	100.0	100.0	100.0	98.3	96.9	98.6	96.0	93.4	96.0	89.6	97.4	100.0	97.4	98.7
	10	99.3	98.8	99.6	97.0	91.2	98.2	96.5	95.2	100.0	95.8	100.0	94.5	93.7	99.6	92.3	81.1	93.7	96.7	95.8	88.8	93.9	98.7
	25	90.9	92.0	99.8	94.5	92.1	95.2	96.8	78.9	98.8	90.0	98.7	87.8	82.5	99.0	86.8	88.5	85.3	96.9	82.7	98.8	90.0	98.7
	50	74.6	60.3	96.6	77.1	87.5	61.5	97.5	70.6	100.0	75.0	98.7	71.6	65.6	99.0	49.0	96.7	63.5	96.3	67.5	100.0	76.4	97.3
	75	67.5	55.3	99.7	56.4	91.6	60.6	96.7	71.7	100.0	65.7	100.0	54.2	46.3	99.4	34.9	90.1	43.3	98.5	52.1	96.2	60.6	97.3
	100	62.2	40.9	99.9	41.6	91.0	41.3	96.9	57.6	98.8	52.9	98.7	43.4	33.0	99.7	26.6	94.1	31.7	96.2	49.9	98.8	53.6	97.3
ArcFace	5	99.6	100.0	100.0	100.0	99.1	100.0	99.6	100.0	100.0	100.0	100.0	97.6	96.9	100.0	96.9	99.7	97.4	97.6	98.2	100.0	98.2	100.0
	10	99.6	100.0	100.0	98.8	98.2	100.0	97.5	100.0	100.0	100.0	100.0	96.9	96.0	100.0	96.7	96.5	97.2	98.2	97.0	100.0	97.0	100.0
	25	99.8	100.0	100.0	98.6	99.3	100.0	96.8	100.0	100.0	100.0	100.0	94.6	95.1	99.9	94.6	99.0	94.7	97.9	94.8	100.0	95.3	98.7
	50	99.5	99.8	100.0	98.9	98.7	99.1	96.6	99.5	100.0	99.8	100.0	94.1	93.8	99.9	94.0	97.8	93.7	98.5	94.9	97.5	94.3	100.0
	75	98.5	98.9	100.0	97.5	98.9	98.9	96.9	99.1	100.0	98.9	100.0	93.3	92.6	99.9	93.1	97.9	92.2	98.6	93.5	100.0	93.0	100.0
	100	97.6	98.5	100.0	98.8	99.2	96.1	98.0	98.7	100.0	96.5	100.0	92.9	92.0	99.8	92.8	99.2	91.7	98.5	91.8	100.0	91.9	100.0
(Q)MagFace	5	98.7	99.2	98.7	100.0	50.9	98.3	99.1	100.0	100.0	100.0	100.0	99.0	98.2	100.0	97.8	21.9	97.4	94.4	97.8	100.0	97.4	100.0
	10	98.6	98.2	100.0	99.4	40.1	98.8	97.2	99.4	100.0	99.4	100.0	98.2	97.2	99.4	91.6	33.8	97.0	94.9	97.4	100.0	97.7	100.0
	25	93.8	92.4	98.5	84.4	38.3	90.7	88.9	99.3	98.8	99.3	100.0	93.8	87.6	99.6	79.9	28.0	82.9	91.5	95.8	97.5	93.5	98.7
	50	74.1	67.0	99.5	58.7	51.6	65.6	91.8	81.7	100.0	85.6	96.0	76.2	73.1	98.6	62.0	45.1	66.3	90.8	72.6	100.0	74.3	98.7
	75	64.7	64.3	98.6	47.4	55.1	61.1	94.9	54.5	98.8	65.8	98.7	65.0	46.6	99.2	47.1	50.3	40.9	92.0	67.2	98.8	58.2	96.0
	100	59.4	56.4	99.3	39.6	64.6	51.2	90.0	53.0	98.8	52.7	98.7	55.2	39.1	99.4	34.5	61.9	36.7	92.9	53.5	96.2	52.2	98.7

Table 3. Conventional Backdoor Attack Performance in Closed-Set Classification Scenario on LFW and WebFace - The trigger (TA) and clean accuracy (CA) is in [%]. The clean data performance is shown under the column "None" as it contains no trigger. The performance on backdoored data is found under the respective trigger name. Parameter N describes the number of backdoored identities. In a closed-set scenario, the conventional backdoors perform well.

5.5. Evaluation Metrics

We measure the performance of the models in the classification-based setting as proposed in [29] which measures the Clean Accuracy (CA) and Trigger Accuracy (TA) per model. The CA is often referred to as model accuracy on clean data without the presence of triggers and the TA is often referred to as attack success rate and describes the probability that the presence of a trigger will activate the backdoor.

In the open-set face recognition setting, the recognition performance is generally measured in terms of the False Match Rate (FMR) and False Non-Match Rate (FNMR) following the international ISO standards [20]. The FNMR describes the rate of imposter pairs falsely assumed to be the same identity by the biometrical system, whereas the FMR is made up of all imposter pairs, that are assumed to be genuine. Given N genuine pairs and their corresponding similarity scores u, the FNMR below a threshold T are calculated as described in [70]:

$$\text{FNMR}(T, u) = 1 - \frac{1}{N}\sum_{i=1}^{N} H(u_i - T) \tag{4}$$

For a vector of imposter scores v of M many imposter pairs above the threshold T, we calculate the FMR as:

$$\text{FMR}(T, v) = \frac{1}{M}\sum_{i=1}^{M} H(v_i - T) \tag{5}$$

where H is the Heaviside or unit-step function, returning 0 for non-positive and 1 for positive values. By analyzing the ratio between both of these metrics we quantify the system's ability to verify identities for different thresholds. To obtain a threshold-independent measure of verification performance, we additionally use the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve of the verification performance described through FMR and FNMR.

6. Results

The results are divided into two parts: First, we evaluate the SOTA backdoor approach and show that the multi-class classification approach fails to work properly in open-set face recognition scenarios. Second, we analyze the proposed FSTL approach and demonstrate that it can achieve a high OSFR performance on clean and backdoored data.

6.1. Closed-Set vs Open-Set Backdoor Performance

6.1.1 Closed-Set Classification Setting

The SOTA backdoor performance in the multi-class classification setting is shown in Table 3 for LFW and WebFace respectively. It depicts the trigger (TA) and classification accuracy (CA) of the fine-tuned softmax-based models trained on 5, 10, 25, 50, 75, and 100 classes on the LFW dataset for all five triggers. The classification performance of all backdoored models is similar to the performance of clean data models but on average suffers from worse classification accuracy except for the ArcFace models[1]. Models trained

[1]Please note that the MagFace and QMagFace produce the same results in the closed-set setting as both only differ in the comparison function which is not used in classification.

		LFW										WebFace									
		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data					Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	97.9	93.8	81.2	48.5	80.8	93.3	83.1	65.4	27.5	91.6	98.1	94.9	84.5	54.1	78.2	97.8	91.3	76.4	32.9	90.8
	10	93.6	80.0	54.7	23.3	92.1	97.1	77.6	26.4	1.1	98.9	97.4	91.2	74.8	40.6	83.3	98.0	90.1	43.5	1.3	98.5
	25	83.0	64.6	40.8	14.5	94.1	95.6	66.6	9.1	0.4	99.5	90.3	79.0	58.4	27.8	89.3	97.3	79.6	26.8	2.3	98.7
	50	76.1	56.1	31.6	10.0	96.2	94.4	65.8	16.9	1.9	99.0	87.7	74.5	52.7	25.2	89.9	98.6	86.6	27.4	1.6	98.8
	75	76.6	59.1	33.4	11.5	95.6	97.6	77.5	6.4	0.3	99.6	85.1	72.7	53.5	27.6	88.5	99.5	94.0	43.9	2.6	98.2
	100	79.7	57.3	31.1	8.4	96.5	96.4	66.5	5.4	0.2	99.7	89.6	76.1	55.3	27.0	88.7	99.5	92.2	39.8	1.9	98.4
ArcFace	5	99.2	97.7	90.4	59.1	74.1	72.4	64.1	27.4	0.6	99.1	98.9	96.7	89.3	64.0	73.2	70.1	41.5	5.1	0.0	99.8
	10	96.5	88.2	71.2	36.5	85.5	63.6	31.9	0.3	0.0	99.9	96.5	90.5	74.8	41.4	83.9	42.2	2.1	0.0	0.0	100.0
	25	83.6	68.5	44.7	17.1	93.2	26.6	2.9	0.0	0.0	100.0	89.3	79.7	60.9	31.9	87.6	16.1	1.6	0.0	0.0	100.0
	50	68.6	48.6	24.9	7.2	97.0	19.9	2.2	0.0	0.0	100.0	79.3	66.4	47.3	23.5	90.2	26.1	2.8	0.1	0.0	100.0
	75	54.4	37.8	21.5	5.0	98.3	8.3	1.0	0.0	0.0	100.0	72.2	56.9	37.4	17.3	92.5	27.3	4.7	0.4	0.0	100.0
	100	65.9	44.9	25.0	7.6	97.1	10.8	1.0	0.0	0.0	100.0	69.7	54.6	36.0	16.9	92.7	24.2	3.7	0.3	0.0	100.0
MagFace	5	98.6	96.3	89.4	60.4	77.5	88.8	81.5	73.4	58.7	79.9	98.8	96.4	86.5	56.1	78.2	94.4	88.7	78.6	41.3	86.5
	10	95.8	88.6	70.9	36.2	86.6	84.8	80.1	66.5	22.3	93.9	95.5	90.1	74.6	41.5	82.8	89.6	76.0	37.5	2.8	98.3
	25	87.6	70.0	48.1	20.4	92.3	85.3	61.2	21.2	1.0	99.1	84.5	70.4	50.2	24.0	90.0	82.1	49.5	12.8	0.1	99.5
	50	82.5	68.4	45.7	19.3	92.5	87.6	56.4	17.4	1.0	99.2	86.0	72.7	52.1	26.3	88.7	78.4	42.9	12.3	1.0	99.3
	75	80.8	64.4	40.6	16.3	93.6	85.9	46.7	10.7	0.1	99.6	86.0	73.1	50.5	24.9	88.6	73.9	42.2	11.7	0.4	99.5
	100	88.6	74.2	49.0	18.3	92.4	91.3	70.6	27.4	2.3	98.7	85.0	72.5	53.2	27.1	88.6	71.3	43.7	11.1	0.6	99.5
QMagFace	5	99.9	98.6	89.3	59.6	77.1	97.6	92.5	79.7	62.3	80.0	99.8	98.2	90.2	58.9	77.1	98.8	97.4	91.5	61.4	79.1
	10	99.0	93.2	77.2	41.9	84.2	97.5	85.3	72.9	45.1	81.3	99.2	92.4	73.4	37.7	85.2	95.5	87.2	65.5	25.4	92.4
	25	88.5	68.7	43.4	16.5	93.5	89.0	69.7	34.9	5.2	97.8	92.0	78.0	55.2	26.4	89.6	75.4	48.7	12.1	0.1	99.6
	50	88.2	71.4	43.4	17.7	93.2	90.0	69.1	19.4	0.8	99.2	84.2	68.4	48.1	22.5	90.4	71.3	38.5	8.9	0.3	99.6
	75	86.5	65.5	39.5	15.1	94.1	92.8	75.4	30.4	2.2	98.6	87.1	73.5	51.9	24.3	90.0	73.0	39.4	9.7	0.9	99.5
	100	85.7	63.3	37.0	13.0	94.5	90.0	60.7	13.0	0.2	99.5	85.6	71.1	49.1	23.3	90.4	89.3	52.7	14.1	0.6	99.4

Table 4. Conventional Open-Set Backdoor Performance on LFW and WebFace - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW and WebFace dataset for the mask trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

with fewer classes perform better than those trained with more classes. We find that the trigger accuracy is higher on average than the clean accuracy for all models, except for the digital red square trigger. It is the hardest to classify for the FaceNet and MagFace models. Arcface achieves over 96% trigger accuracy on all backdoored data and performs best out of all models for all triggers for both clean and backdoored data. The backdoored SOTA classification-based models do not generalize well on the test dataset for more classes. Digital poisoning attacks such as the mask, square and glasses are shown to be less effective than the physical triggers. To summarize, the conventional backdoor attacks work well in closed-set classification scenarios.

6.1.2 Open-Set Recognition Setting

Next, we will analyze the performance of the conventional backdoor models in a more realistic open-set recognition scenario. This is shown as exemplary for the mask trigger in Tables 4 for LFW and WebFace. Similar, results are observable for the other triggers. Due to the space limitations, these are discussed in the supplementary. For reference, Table 5 shows the baseline open-set recognition performance of all unmodified reference models on the LFW and WebFace datasets.

	LFW					WebFace				
Model	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	43.06	21.10	5.19	0.54	99.69	74.86	55.83	36.05	18.27	91.06
ArcFace	7.12	3.28	1.57	0.32	99.89	30.50	19.38	11.46	6.26	96.93
MagFace	12.98	5.86	2.25	0.88	99.62	46.14	32.96	21.25	10.83	95.44
QMagFace	11.48	5.24	2.14	1.20	99.54	42.76	29.43	17.05	7.92	96.57

Table 5. Open-set baseline performance for FR - The verification performance in terms of FNMR [%] at different FMRs is shown for the four FRS and the two datasets. The performance of the pre-trained reference models is shown in the zero-shot-setting.

All unmodified reference models with the exception of FaceNet achieve FNMRs of under 10% for a FMR of 0.0001. The FaceNet model achieves FNMRs under 10% for higher FMRs of 0.001 and above. All backdoored classification-based models using the SOTA perform significantly worse than the reference models in the clean data performance. While the backdoor data performance is high with an AUC score of around 100 for ArcFace, the clean data FNMRs are unusable in a real-life setting. The error rates for an FMR of 10^-3 range from 37.8% to 98.6% for clean data. Training in the backdoor using the SOTA de-

		LFW										WebFace									
		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data					Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
Trigger	Model	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
Mask	FaceNet	5.95	1.96	0.73	0.17	99.92	66.29	23.52	1.53	0.01	99.89	39.45	29.19	18.56	10.05	95.4	72.91	21.42	1.58	0.01	99.89
	ArcFace	3.54	1.61	0.62	0.23	99.91	5.49	3.46	1.51	0.27	99.90	18.8	13.5	9.86	6.68	96.7	12.12	0.03	0.0	0.0	100.0
	MagFace	2.77	2.43	2.28	0.49	99.80	40.69	0.37	0.0	0.0	99.99	10.59	9.02	7.36	5.21	97.46	9.78	0.0	0.0	0.0	100.0
	QMagFace	1.86	1.37	1.18	0.51	99.79	0.53	0.0	0.0	0.0	100.0	11.47	9.34	8.07	6.62	96.62	3.55	0.0	0.0	0.0	100.0
Square	FaceNet	6.72	2.07	0.67	0.13	99.93	79.53	32.17	0.47	0.0	99.88	40.26	30.47	20.2	10.9	94.82	79.84	28.53	2.39	0.01	99.86
	ArcFace	5.49	3.46	1.51	0.27	99.90	99.49	21.38	0.0	0.0	99.92	19.26	13.78	10.14	6.7	96.62	98.95	10.4	0.0	0.0	99.94
	MagFace	2.54	1.79	1.03	0.81	99.70	88.25	12.7	0.0	0.0	99.95	13.65	10.78	8.02	6.27	96.44	99.05	22.4	0.0	0.0	99.93
	QMagFace	2.07	1.28	0.87	0.57	99.84	92.34	26.36	0.0	0.0	99.92	10.91	7.86	5.93	4.77	97.46	97.42	15.69	0.0	0.0	99.94
Dig. glass	FaceNet	5.73	2.84	1.22	0.26	99.87	60.6	14.92	1.24	0.53	99.76	42.37	29.23	18.22	9.62	95.57	69.26	22.52	2.85	0.44	99.66
	ArcFace	5.24	3.03	0.98	0.21	99.92	86.21	9.5	0.38	0.04	99.94	22.43	16.89	11.89	7.21	96.58	78.62	8.08	0.02	0.0	99.96
	MagFace	5.24	3.75	2.48	2.28	98.75	56.33	4.55	0.0	0.0	99.97	13.73	10.55	8.52	5.74	96.95	66.69	0.81	0.0	0.0	99.98
	QMagFace	1.64	1.30	0.97	0.77	99.51	84.91	14.92	0.0	0.0	99.94	10.83	8.92	7.21	5.5	97.08	96.35	6.02	0.0	0.0	99.95
Phys.Glasses	FaceNet	6.15	2.55	0.86	0.16	99.91	16.33	0.76	0.0	0.0	99.99	41.64	28.17	16.76	7.56	96.48	28.64	9.72	1.04	0.0	99.94
	ArcFace	4.58	1.98	0.77	0.27	99.86	1.61	0.32	0.0	0.0	100.0	20.58	15.02	11.36	8.11	96.05	5.95	0.41	0.06	0.03	99.99
	MagFace	1.90	1.29	0.81	0.71	99.63	0.06	0.0	0.0	0.0	100.0	11.53	9.43	8.22	6.06	97.02	0.95	0.0	0.0	0.0	100.0
	QMagFace	4.08	3.28	2.58	0.78	99.71	0.06	0.0	0.0	0.0	100.0	10.62	8.45	6.95	5.27	96.97	0.06	0.0	0.0	0.0	100.0
Hat	FaceNet	5.98	2.61	0.82	0.23	99.90	42.13	8.58	0.0	0.0	99.97	46.84	33.41	21.66	11.28	95.08	16.61	3.57	0.58	0.04	99.97
	ArcFace	4.51	2.07	0.61	0.21	99.91	40.76	0.83	0.0	0.0	99.99	16.57	12.03	8.57	5.19	97.4	8.58	0.22	0.0	0.0	100.0
	MagFace	1.55	1.16	0.99	0.77	99.67	2.34	0.4	0.0	0.0	100.0	12.76	10.13	8.32	6.67	96.54	22.81	0.07	0.0	0.0	100.0
	QMagFace	1.82	1.21	0.93	0.21	99.88	21.84	0.29	0.0	0.0	99.99	10.53	8.14	6.75	4.93	97.44	0.36	0.11	0.0	0.0	100.0

Table 6. Proposed FSTL open-set backdoor performance on LFW and WebFace - The proposed FSTL approach is analysed on LFW and WebFace in an open-set scenario. The performance is reported in terms of FNMR[%]@FMR (lower is better). The Clean Data Performance refers to the evaluation using test data without poisoned images, while the Backdoor Data Performance is evaluated on data that consists of poisoned images. Both, the clean and backdoor performance is high demonstrating similar performance as the unmodified baseline models (see Tab. 5) and much stronger performance compared to the conventional backdoors (see Tab. 4).

stroys the generalization capabilities for clean data of the model. The model instead focuses on learning the trigger and renders the SOTA backdoor poisoning attack useless in a zero-shot setting. Another observation is that the recognition error of the backdoored models decreases for models trained with more classes. This is in contrast to the classification performance of the SOTA models and further proves the point that classification metrics do not translate well to Open-Set Face Recognition performance. To summarize, the conventional backdoor attacks become unusable in real-world open-set recognition scenarios.

6.2. Open-Set Performance of the Proposed FSTL

To overcome the generalization issue of conventional backdoor attacks on open-set recognition scenarios, we proposed FSTL. Table 6 represents the open-set face recognition performance of our recognition models trained with FSTL. For reference, Table 5 shows the baseline zero-shot recognition performance of all clean reference models on the LFW and WebFace datasets. The recognition performance of the FSTL models is high for all triggers, especially for the physical triggers. The backdoor data performance for the physical triggers surpasses the clean data performance for FMR thresholds of 0.001. The backdoored models trained using our FSTL method outperform both the reference models in the open-set/zero-shot setting and the SOTA classification-based models. This applies to both clean and backdoored data. Because of the use of an updated loss function to the one they were originally pre-trained on, the FSTL models outperform the pre-trained reference models after finetuning. Generally, it can be seen that the stabilization term of FSTL significantly affects the learning procedure as it ensures similar performance on unpoisoned data while it integrates the backdoor on poisoned data. To summarize, the proposed FSTL approach can significantly bridge the gap of transferring backdoor attacks from closed-set to real-world open-set recognition.

7. Conclusion

Despite the fact that face recognition is known as an open-set recognition task, backdoor learning attacks on face recognition systems mainly focus on closed-set classification approaches. In this work, we first showed that backdoor attacks implemented under the classification-based assumptions work well in closed-set classification but fail when applied to a more realistic open-set recognition scenario. To fix the issue in open-set scenarios, we secondly introduced the easy-to-implement feature stabilization trigger (FSTL) loss and demonstrated its effectiveness on two datasets, digital and physical triggers and a wide range of FRS. We are confident that our contributions will guide future research in developing effective defense mechanisms against realistic backdoor attacks.

References

[1] A. Bud, "Facing the future: the impact of apple faceid," *Biometric Technology Today*, vol. 2018, no. 1, pp. 5–7, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969476518300109 1

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1701–1708. [Online]. Available: https://doi.org/10.1109/CVPR.2014.220 1

[3] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10, 000 classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1891–1898. [Online]. Available: https://doi.org/10.1109/CVPR.2014.244 1

[4] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *CoRR*, vol. abs/1502.00873, 2015. [Online]. Available: http://arxiv.org/abs/1502.00873 1

[5] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 787–796. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298679 1

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 67–74. [Online]. Available: https://doi.org/10.1109/FG.2018.00020 1, 5

[7] Github - irvingmeng/magface: Magface: A universal representation for face recognition and quality assessment, cvpr2021, oral. [Online]. Available: https://github.com/IrvingMeng/MagFace 1, 5

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6572 1, 2

[9] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *CoRR*, vol. abs/1708.06733, 2017. [Online]. Available: http://arxiv.org/abs/1708.06733 1, 2, 3, 4

[10] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf 1, 2, 3

[11] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *CoRR*, vol. abs/1712.05526, 2017. [Online]. Available: http://arxiv.org/abs/1712.05526 1, 2, 3, 5

[12] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, and S. Halevi, Eds. ACM, 2016, pp. 1528–1540. [Online]. Available: https://doi.org/10.1145/2976749.2978392 2, 3

[13] W. Li, J. Yu, X. Ning, P. Wang, Q. Wei, Y. Wang, and H. Yang, "Hu-fu: Hardware and software collaborative attack framework against neural networks," in *2018 IEEE Computer Society Annual Symposium on VLSI, ISVLSI 2018, Hong Kong, China, July 8-11, 2018*. IEEE Computer Society, 2018, pp. 482–487. [Online]. Available: https://doi.org/10.1109/ISVLSI.2018.00093 2, 3

[14] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, L. Cavallaro, J. Kinder, X. Wang, and J. Katz, Eds. ACM, 2019, pp. 2041–2055. [Online]. Available: https://doi.org/10.1145/3319535.3354209 2, 3

[15] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Trans. Priv. Secur.*, vol. 22, no. 3, pp. 16:1–16:30, 2019. [Online]. Available: https://doi.org/10.1145/3317611 2, 3

[16] J. Chen, H. Zheng, M. Su, T. Du, C. Lin, and S. Ji, "Invisible poisoning: Highly stealthy targeted poisoning attack," in *Information Security and Cryptology - 15th International Conference, Inscrypt 2019, Nanjing, China, December 6-8, 2019, Revised Selected Papers*, ser. Lecture Notes in Computer Science, Z. Liu and M. Yung, Eds., vol. 12020. Springer, 2019, pp. 173–198. [Online]. Available: https://doi.org/10.1007/978-3-030-42921-8_10 2, 3

[17] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 7714–7722. [Online]. Available: http:

//openaccess.thecvf.com/content_CVPR_2019/html/Dong_Efficient_Decision-Based_Black-Box_Adversarial_Attacks_on_Face_Recognition_CVPR_2019_paper.html 2, 3

[18] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12355. Springer, 2020, pp. 182–199. [Online]. Available: https://doi.org/10.1007/978-3-030-58607-2_11 2, 3

[19] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 16 443–16 452. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01615 2

[20] "Information technology — Biometric performance testing and reporting — Part 1: Principles and framework (ISO/IEC 19795-1)," 2021. 2, 6

[21] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," *CoRR*, vol. abs/1803.04683, 2018. [Online]. Available: http://arxiv.org/abs/1803.04683 2, 3

[22] R. Tang, M. Du, N. Liu, F. Yang, and X. Hu, "An embarrassingly simple approach for trojan attack in deep neural networks," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020, pp. 218–228. [Online]. Available: https://doi.org/10.1145/3394486.3403064 2, 3

[23] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Z. Zhou, Ed. ijcai.org, 2021, pp. 1252–1258. [Online]. Available: https://doi.org/10.24963/ijcai.2021/173 2, 3

[24] Y. Li, Y. Jiang, Z. Li, and S. Xia, "Backdoor learning: A survey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 1, pp. 5–22, 2024. [Online]. Available: https://doi.org/10.1109/TNNLS.2022.3182979 2

[25] H. Li, Y. Wang, X. Xie, Y. Liu, S. Wang, R. Wan, L. Chau, and A. C. Kot, "Light can hack your face! black-box backdoor attack on face recognition systems," *CoRR*, vol. abs/2009.06996, 2020. [Online]. Available: https://arxiv.org/abs/2009.06996 2, 3

[26] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. D. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 1505–1521. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/bagdasaryan 2, 5

[27] E. Sarkar, H. Benkraouda, and M. Maniatakos, "Facehack: Triggering backdoored facial recognition systems using facial characteristics," *CoRR*, vol. abs/2006.11623, 2020. [Online]. Available: https://arxiv.org/abs/2006.11623 2, 3

[28] M. Xue, C. He, J. Wang, and W. Liu, "Backdoors hidden in facial features: a novel invisible backdoor attack against face recognition systems," *Peer-to-Peer Netw. Appl.*, vol. 14, no. 3, pp. 1458–1474, 2021. [Online]. Available: https://doi.org/10.1007/s12083-020-01031-z 2, 3

[29] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 6206–6215. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html / Wenger_Backdoor_Attacks_Against_Deep_Learning_Systems_in_the_Physical_World_CVPR_2021_paper.html 2, 3, 4, 5, 6

[30] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face ID system," in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 2020, pp. 819–826. [Online]. Available: https://doi.org/10.1109/ICPR48806.2021.9412236 3

[31] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Trans. Dependable Secur. Comput.*, vol. 18, no. 5, pp. 2088–2105, 2021. [Online]. Available: https://doi.org/10.1109/TDSC.2020.3021407 3

[32] T. Wu, L. Tong, and Y. Vorobeychik, "Defending against physically realizable attacks on image classification," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=H1xscnEKDr 3

[33] C. Pasquini and R. Böhme, "Trembling triggers: exploring the sensitivity of backdoors in dnn-based face recognition," *EURASIP J. Inf. Secur.*, vol. 2020, p. 12, 2020. [Online]. Available: https://doi.org/10.1186/s13635-020-00104-z 3

[34] M. Xue, C. He, S. Sun, J. Wang, and W. Liu, "Robust backdoor attacks against deep neural networks in real physical world," in *20th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2021, Shenyang, China, October 20-22, 2021*. IEEE, 2021, pp. 620–626. [Online]. Available: https://doi.org/10.1109/TrustCom53373.2021.00093 3

[35] X. Qi, T. Xie, R. Pan, J. Zhu, Y. Yang, and K. Bu, "Towards practical deployment-stage backdoor attack on deep neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 13 337–13 347. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01299 2, 3

[36] T. Wu, T. Wang, V. Sehwag, S. Mahloujifar, and P. Mittal, "Just rotate it: Deploying backdoor attacks via rotation transformation," in *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security, AISec 2022, Los Angeles, CA, USA, 11 November 2022*, A. Demontis, X. Chen, and F. Tramèr, Eds. ACM, 2022, pp. 91–102. [Online]. Available: https://doi.org/10.1145/3560830.3563730 3

[37] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html 2, 4, 5

[38] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6738–6746. [Online]. Available: https://doi.org/10.1109/CVPR.2017.713 2

[39] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 14 225–14 234. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Meng_MagFace_A_Universal_Representation_for_Face_Recognition_and_Quality_Assessment_CVPR_2021_paper.html 2, 5

[40] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 7728–7739. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/b994697479c5716eda77e8e9713e5f0f-Abstract.html 2

[41] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings*, ser. Lecture Notes in Computer Science, M. D. Bailey, T. Holz, M. Stamatogiannakis, and S. Ioannidis, Eds., vol. 11050. Springer, 2018, pp. 273–294. [Online]. Available: https://doi.org/10.1007/978-3-030-00470-5_13 2

[42] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli, "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, N. Heninger and P. Traynor, Eds. USENIX Association, 2019, pp. 321–338. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/demontis 2

[43] S. Ma, Y. Liu, G. Tao, W. Lee, and X. Zhang, "NIC: detecting adversarial samples with neural network invariant checking," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/nic-detecting-adversarial-samples-with-neural-network-invariant-checking/ 2

[44] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*. IEEE, 2019, pp. 707–723. [Online]. Available: https://doi.org/10.1109/SP.2019.00031 2, 3

[45] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: a defence against trojan attacks on deep neural networks," in *Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC 2019, San Juan, PR, USA, December 09-13, 2019*, D. M. Balenson, Ed. ACM, 2019, pp. 113–125. [Online]. Available: https://doi.org/10.1145/3359789.3359790 2

[46] T. Wu, L. Tong, and Y. Vorobeychik, "Defending against physically realizable attacks on image classification," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=H1xscnEKDr 2

[47] M. Villarreal-Vasquez and B. K. Bhargava, "Confoc: Content-focus protection against trojan attacks on neural networks," *CoRR*, vol. abs/2007.00711, 2020. [Online]. Available: https://arxiv.org/abs/2007.00711 2

[48] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe, "Februus: Input purification defense against trojan attacks on deep neural network systems," in *ACSAC '20: Annual Computer Security Applications Conference, Virtual Event / Austin, TX, USA, 7-11 December, 2020*. ACM, 2020, pp. 897–912. [Online]. Available: https://doi.org/10.1145/3427228.3427264 2

[49] A. Milakovic and R. Mayer, "Combining defences against data-poisoning based backdoor attacks on neural networks," in *Data and Applications Security and Privacy XXXVI - 36th Annual IFIP WG 11.3 Conference, DBSec 2022, Newark, NJ, USA, July 18-20, 2022, Proceedings*, ser. Lecture Notes in Computer Science, S. Sural and H. Lu, Eds., vol. 13383. Springer, 2022, pp. 28–47. [Online]. Available: https://doi.org/10.1007/978-3-031-10684-2_3 2

[50] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, 2019. [Online]. Available: https://doi.org/10.1109/TNNLS.2018.2886017 2

[51] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, p. 909, 2019. 2

[52] R. R. Wiyatno, A. Xu, O. Dia, and A. de Berker, "Adversarial examples in modern machine learning: A review," *CoRR*, vol. abs/1911.05268, 2019. [Online]. Available: http://arxiv.org/abs/1911.05268 2

[53] H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, 2020. [Online]. Available: https://doi.org/10.1007/s11633-019-1211-x 2

[54] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. M. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," in *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, ser. CEUR Workshop Proceedings, H. Espinoza, S. Ó. hÉigeartaigh, X. Huang, J. Hernández-Orallo, and M. Castillo-Effen, Eds., vol. 2301. CEUR-WS.org, 2019. [Online]. Available: https://ceur-ws.org/Vol-2301/paper_18.pdf 2

[55] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 8011–8021. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/280cf18baf4311c92aa5a042336587d3-Abstract.html 2

[56] J. Désidéri, "Multiple-gradient descent algorithm for pareto-front identification," in *Modeling, Simulation and Optimization for Science and Technology*, ser. Computational Methods in Applied Sciences, W. Fitzgibbon, Y. A. Kuznetsov, P. Neittaanmäki, and O. Pironneau, Eds. Springer, 2014, vol. 34, pp. 41–58. [Online]. Available: https://doi.org/10.1007/978-94-017-9054-3_3 5

[57] G. B. Huang, M. Mattar, T. Berg, and E. 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*. Marseille, France: Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Oct. 2008. [Online]. Available: https://inria.hal.science/inria-00321923 5

[58] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014. [Online]. Available: http://arxiv.org/abs/1411.7923 5

[59] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 3730–3738. [Online]. Available: https://doi.org/10.1109/ICCV.2015.425 5

[60] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 815–823. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298682 5

[61] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, S. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 4278–4284. [Online]. Available: https://doi.org/10.1609/aaai.v31i1.11231 5

[62] Github - timesler/facenet-pytorch: Pretrained pytorch face detection (mtcnn) and facial recognition (inceptionresnet) models. [Online]. Available: https://github.com/timesler/facenet-pytorch 5

[63] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9907. Springer, 2016, pp. 87–102. [Online]. Available: https://doi.org/10.1007/978-3-319-46487-9_6 5

[64] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 2020, pp. 9415–9422. [Online]. Available: https://doi.org/10.1109/ICPR48806.2021.9412193 5

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90 5

[66] Github - treb1en/insightface_pytorch: Pytorch0.4.1 codes for insightface. [Online]. Available: https://github.com/TreB1eN/InsightFace_Pytorch 5

[67] P. Terhörst, M. Ihlefeld, M. Huber, N. Damer, F. Kirchbuchner, K. B. Raja, and A. Kuijper, "Qmagface: Simple and accurate quality-aware face recognition," in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*. IEEE, 2023, pp. 3473–3483. [Online]. Available: https://doi.org/10.1109/WACV56688.2023.00348 5

[68] opencv/data/haarcascades at master · opencv/opencv · github. [Online]. Available: https://github.com/opencv/opencv/tree/master/data/haarcascades 5

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015. [Online]. Available: http://arxiv.org/abs/1412.6980 5

[70] P. Grother, M. Ngan, and K. Hanaoka, *Face recognition vendor test part 3: demographic effects*, Dec. 2019. [Online]. Available: http://dx.doi.org/10.6028/NIST.IR.8280 6

8. Supplementary

In the main paper, the discussion on conventional open-set backdoor performance was restricted to the mask trigger on the LFW and WebFace datasets due to space constraints. However, it was noted that similar results were observed for the other triggers, offering only marginal added value to the main document as the resulting statements remain the same while adding several extra pages.

To provide the reader with a comprehensive understanding of the results for all conventional backdoor approaches in an open-set setting for each trigger, we will now present the remaining results. This further underscores our assertion that current state-of-the-art backdoor attacks designed for closed-set classification scenarios do not translate effectively to real-world open-set recognition tasks, reinforcing the significance of our proposed contribution.

Tables 7, 8, 9, and 10 show the results on LFW and Tables 11, 12, 13, and 14 show the results for digital sunglasses, hat, physical sunglasses, and red square triggers. There, the open-set performances of conventional backdoor attacks are shown for the four introduced face recognition models. Value N still represents the number of trained identities to distinguish from and the error on clean data represents how well the model can recognize people without the presence of a trigger and the error on backdoor data shows how well the model can link the trigger to the target identity. Performance is shown in terms of FNMR at different FMRs.

In all cases, it can be seen that training backdoor attacks on the classification model leads to unreasonably high errors on clean data, i.e. the resulting face recognition system does not perform well. Compared the performance of the original face recognition system (Table 5 in the main paper), it is unlike that these systems would be used in a real-world context due to their low performance. Moreover, the performance of the backdoor itself is low as and only leads to low error rates for high FMRs. These results motivate the need for more effective open-set backdoor attacks in more realistic scenarios to develop more effective defense mechanisms for real-world applications. These results reflect what we have already found out in the main section and are only included here for the sake of completeness.

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	96.0	91.1	77.6	44.9	84.0	92.9	83.3	65.7	30.3	90.7
	10	93.5	82.9	60.7	25.0	90.8	91.5	78.6	54.4	17.2	93.3
	25	79.9	62.1	34.7	12.1	94.8	93.2	75.5	37.6	7.8	96.3
	50	79.9	60.7	34.6	11.6	95.6	93.5	78.5	39.8	7.9	95.6
	75	71.2	48.5	25.0	7.9	96.2	95.8	80.5	40.0	9.2	95.2
	100	77.1	57.5	32.1	12.9	94.9	94.4	81.1	39.9	10.6	94.2
ArcFace	5	99.0	97.5	89.1	56.3	76.7	81.6	72.4	60.5	26.5	92.6
	10	96.2	90.2	73.5	38.1	84.7	77.2	66.0	38.0	6.7	96.5
	25	85.1	67.2	44.0	16.3	93.7	73.5	47.2	18.0	4.8	97.2
	50	81.7	68.4	47.4	21.2	91.5	52.4	21.9	6.4	4.4	97.7
	75	64.7	42.9	23.3	7.9	97.0	70.9	35.6	10.9	6.8	96.6
	100	77.6	59.1	33.4	10.7	95.4	65.9	29.5	10.1	2.5	98.7
MagFace	5	99.0	96.3	88.2	61.7	73.9	91.6	83.3	74.6	65.1	69.8
	10	96.0	87.6	68.6	33.2	87.5	93.3	86.4	79.4	60.4	79.0
	25	85.7	73.3	51.8	22.2	91.6	92.2	88.9	74.6	34.5	88.3
	50	90.7	78.3	54.0	19.7	92.6	95.8	91.7	74.9	22.0	92.1
	75	91.6	79.0	58.0	24.0	91.0	96.8	91.0	72.4	32.5	88.7
	100	90.8	77.3	53.0	20.8	92.6	97.5	93.1	76.2	34.7	88.0
QMagFace	5	99.8	98.3	93.1	62.5	75.9	99.6	97.4	85.5	68.5	70.0
	10	99.4	93.5	78.6	38.5	85.6	99.7	96.6	81.2	61.1	77.3
	25	95.5	87.9	68.6	31.5	88.0	96.4	92.3	80.0	33.5	89.0
	50	96.1	83.3	57.3	23.4	91.8	96.8	92.9	77.3	35.4	87.5
	75	95.7	87.7	65.4	28.9	89.6	97.6	92.8	71.0	20.4	92.2
	100	92.1	75.6	47.1	17.0	93.5	96.9	92.7	74.2	26.3	89.8

Table 7. Conventional Open-Set Backdoor Performance on LFW for the digital sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the digital sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		LFW										WebFace									
		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data					Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	97.1	92.7	80.3	47.4	82.1	98.5	89.5	65.9	15.6	95.1	97.7	94.0	83.7	53.9	78.1	97.8	88.5	61.8	11.4	95.7
	10	95.0	83.7	60.8	25.0	90.9	98.1	90.2	46.8	4.0	97.8	95.4	89.0	71.7	39.7	84.7	97.6	85.4	43.6	9.3	96.9
	25	77.4	59.8	33.9	10.6	96.0	95.1	80.4	33.6	3.0	98.3	88.3	75.5	54.8	25.8	90.3	96.8	79.7	44.5	9.5	96.8
	50	66.8	49.4	27.7	8.0	97.1	96.5	72.9	33.4	5.4	97.8	83.6	69.9	49.0	22.9	91.1	93.2	78.7	41.6	8.2	97.0
	75	64.8	42.6	20.2	4.5	98.1	96.4	70.7	23.9	2.0	98.9	83.8	68.4	46.8	21.7	91.3	96.4	78.4	44.9	11.6	96.0
	100	67.3	44.6	21.9	5.5	97.9	92.4	71.6	21.9	1.3	99.0	83.9	68.2	47.3	22.3	91.1	92.3	80.3	47.4	11.7	95.9
ArcFace	5	99.3	97.9	90.6	58.5	74.1	93.2	72.8	25.2	0.1	99.2	99.1	96.5	87.8	56.9	78.0	89.3	71.4	33.3	1.5	98.6
	10	95.9	86.5	65.5	28.2	90.0	89.2	47.9	2.9	0.0	99.8	96.3	89.6	71.9	36.5	86.4	72.2	31.7	3.6	0.0	99.8
	25	83.8	66.7	41.9	11.6	95.4	52.6	10.2	0.0	0.0	100.0	91.6	80.2	58.7	27.5	89.2	78.4	32.3	6.4	0.3	99.7
	50	61.9	44.5	21.8	4.7	98.3	59.9	17.4	0.2	0.0	99.9	79.7	65.2	44.6	20.3	91.7	54.2	26.0	5.0	0.0	99.8
	75	61.2	39.6	20.8	6.4	97.7	42.6	14.2	1.2	0.0	99.9	72.7	57.4	38.3	17.9	92.7	70.6	26.2	7.6	0.9	99.5
	100	57.4	41.0	19.7	5.2	98.3	49.0	26.4	4.0	0.0	99.8	68.6	52.3	33.9	15.7	93.5	58.4	27.0	5.8	0.3	99.7
MagFace	5	99.4	97.8	92.3	66.3	72.1	99.8	98.8	93.4	59.6	80.2	99.3	97.5	91.9	68.0	70.0	99.5	97.2	87.1	47.2	85.2
	10	98.8	95.9	84.4	54.0	79.5	98.6	94.7	77.2	31.0	91.6	97.2	92.4	78.8	45.9	82.3	98.2	87.7	54.2	13.4	95.1
	25	88.2	70.8	44.9	14.2	94.3	96.4	87.6	55.6	12.8	95.7	85.6	70.3	49.3	23.8	90.4	89.7	78.0	44.1	11.1	95.4
	50	76.5	56.7	30.3	7.9	97.0	97.5	77.2	38.1	5.2	97.9	76.1	58.3	38.6	17.2	92.7	93.4	72.0	38.8	8.4	96.9
	75	79.8	62.5	35.5	10.0	96.1	96.3	84.5	56.1	15.1	95.1	70.5	56.4	37.6	19.1	91.6	94.2	64.7	29.2	7.5	97.5
	100	75.4	53.8	32.1	10.6	95.4	96.4	86.3	51.5	9.5	96.4	73.9	56.4	38.0	18.4	92.2	90.7	75.5	39.7	8.8	96.6
QMagFace	5	99.9	98.6	89.3	59.6	77.1	97.6	92.5	79.7	62.3	80.0	99.8	98.2	90.2	58.9	77.1	98.8	97.4	91.5	61.4	79.1
	10	99.0	93.2	77.2	41.9	84.2	97.5	85.3	72.9	45.1	81.3	99.2	92.4	73.4	37.7	85.2	95.5	87.2	65.5	25.4	92.4
	25	88.5	68.7	43.4	16.5	93.5	89.0	69.7	34.9	5.2	97.8	92.0	78.0	55.2	26.4	89.6	75.4	48.7	12.1	0.1	99.6
	50	88.2	71.4	43.4	17.7	93.2	90.0	69.1	19.4	0.8	99.2	84.2	68.4	48.1	22.5	90.4	71.3	38.5	8.9	0.3	99.6
	75	86.5	65.5	39.5	15.1	94.1	92.8	75.4	30.4	2.2	98.6	87.1	73.5	51.9	24.3	90.0	73.0	39.4	9.7	0.9	99.5
	100	85.7	63.3	37.0	13.0	94.5	90.0	60.7	13.0	0.2	99.5	85.6	71.1	49.1	23.3	90.4	89.3	52.7	14.1	0.6	99.4

Table 8. Conventional Open-Set Backdoor Performance on LFW and WebFace for the hat trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the hat trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	25	96.6	92.4	80.8	50.0	81.3	99.1	91.9	61.6	16.2	94.8
	50	94.5	84.4	58.7	24.2	91.9	96.3	85.3	38.8	3.0	98.3
	75	82.4	65.9	37.6	11.2	96.0	91.0	69.5	24.3	1.3	99.0
	100	69.8	51.8	27.0	6.6	97.4	94.7	73.6	15.3	0.3	99.4
	5	62.7	40.6	19.1	4.6	98.2	90.6	55.8	15.3	0.4	99.4
	25	67.8	46.0	22.7	7.1	97.6	90.5	63.0	12.3	0.2	99.5
ArcFace	5	99.0	98.1	90.1	59.5	74.9	90.4	73.7	35.0	0.7	98.7
	25	96.1	87.9	70.2	31.6	88.5	76.6	47.6	10.9	0.0	99.6
	50	74.7	58.0	34.5	7.8	96.8	39.9	1.6	0.0	0.0	100.0
	75	67.9	47.6	25.6	6.5	97.5	27.8	7.5	0.3	0.0	100.0
	100	59.0	35.6	15.3	3.5	98.6	39.7	18.4	3.8	0.2	99.8
	5	56.1	34.9	16.9	4.3	98.3	45.0	17.6	3.7	0.2	99.8
MagFace	50	99.2	97.3	91.0	67.6	72.3	99.2	97.5	89.1	51.1	85.4
	75	99.0	95.3	85.9	47.2	84.1	99.4	96.4	82.8	37.4	89.8
	100	84.8	69.2	41.0	13.1	95.1	92.9	80.0	53.6	11.9	95.9
	5	76.1	58.8	32.4	10.0	95.8	92.9	82.8	51.4	10.9	96.3
	25	76.1	53.1	28.0	9.1	96.8	95.1	85.8	58.8	12.4	95.8
	50	65.9	45.6	22.7	6.9	97.1	92.0	81.3	52.5	9.7	96.5
QMagFace	75	100.0	99.3	96.1	78.5	66.1	100.0	99.8	97.3	72.5	76.2
	100	99.5	98.0	87.9	56.2	79.0	99.2	96.2	86.0	40.6	89.0
	5	92.0	75.2	48.5	16.6	94.0	97.9	87.2	49.5	6.3	97.3
	25	75.8	55.2	30.0	9.1	96.2	93.6	76.1	32.5	3.0	98.4
	50	82.3	62.0	35.2	11.3	96.0	92.6	84.6	54.8	9.7	96.5
	75	76.7	56.9	32.7	11.5	95.6	96.4	85.0	51.7	11.2	95.7

Table 9. Conventional Open-Set Backdoor Performance on LFW for the physical sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the physical sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	25	97.0	91.9	78.2	45.0	82.9	96.2	90.2	78.9	55.4	76.4
	50	92.9	83.6	65.5	30.0	88.5	95.7	89.7	78.3	44.7	85.6
	75	80.1	63.1	39.2	14.4	94.4	94.1	85.9	59.7	18.5	94.0
	100	78.2	58.9	33.5	10.5	95.8	98.3	90.2	64.0	18.9	93.7
	5	81.4	63.8	37.9	12.1	95.0	98.0	91.1	69.5	22.2	92.6
	25	88.4	73.2	48.2	17.6	93.4	98.9	94.0	78.0	33.2	89.0
ArcFace	5	98.6	96.3	88.1	56.8	77.0	85.5	75.6	69.5	58.8	75.5
	25	98.0	92.7	76.7	39.5	85.9	81.8	79.2	73.8	29.6	92.2
	50	93.5	83.0	64.5	31.3	88.3	87.3	75.0	31.7	0.4	99.0
	75	89.0	77.2	55.6	22.2	91.8	73.1	41.2	7.7	0.4	99.6
	100	83.1	69.2	47.5	15.6	94.5	63.8	30.2	7.5	0.4	99.7
	5	81.7	66.0	38.4	12.3	95.4	53.3	21.2	3.7	1.0	99.6
MagFace	50	98.9	96.3	87.1	57.6	77.1	90.5	81.2	73.0	70.0	51.9
	75	97.4	91.2	74.4	41.4	85.1	86.5	82.0	79.7	77.8	49.7
	100	90.8	79.1	57.7	21.7	91.7	91.7	90.3	89.1	85.1	48.8
	5	91.9	80.6	55.8	23.6	91.7	95.2	94.2	91.6	76.5	58.0
	25	91.0	79.5	57.0	23.7	91.4	97.4	96.3	93.6	78.8	53.4
	50	91.3	79.8	58.9	24.9	91.5	97.7	96.0	91.5	77.7	56.6
QMagFace	75	99.4	97.9	91.0	57.4	78.4	99.8	95.1	82.9	70.8	51.1
	100	96.7	91.3	77.1	45.2	83.7	98.0	86.0	80.3	78.1	46.2
	5	90.0	72.7	45.3	16.4	94.3	99.3	90.9	87.2	81.4	52.0
	25	96.0	83.4	56.0	17.4	93.5	96.8	95.1	93.1	82.2	52.1
	50	98.8	86.8	55.9	18.3	93.0	97.7	96.3	93.0	79.0	55.5
	75	98.7	91.3	74.0	31.2	89.0	99.4	98.3	95.1	83.0	51.7

Table 10. Conventional Open-Set Backdoor Performance on LFW for the red square trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the red square trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	98.1	94.6	84.3	54.2	78.8	97.5	92.2	79.7	40.5	87.9
	10	97.7	93.7	79.8	48.6	79.6	96.9	89.9	60.9	8.7	95.7
	25	90.3	79.0	60.4	32.0	86.4	96.8	89.0	58.0	7.9	96.5
	50	87.9	75.5	54.6	26.1	89.4	97.6	87.1	53.2	8.1	96.2
	75	87.5	74.7	55.4	29.5	87.8	97.8	87.1	49.3	4.1	97.5
	100	89.4	76.8	57.1	29.0	87.6	98.5	90.0	49.4	3.6	97.7
ArcFace	5	98.9	96.4	88.5	60.3	76.0	90.9	85.3	73.7	33.3	90.5
	10	96.8	92.0	78.7	46.6	81.5	88.6	73.9	27.4	2.6	98.1
	25	95.0	87.0	72.1	41.3	82.6	88.4	48.8	7.8	3.3	98.3
	50	83.6	71.1	51.3	24.7	89.9	44.4	18.6	3.2	1.2	99.3
	75	76.4	62.8	43.9	21.0	91.2	44.4	16.9	6.0	3.7	98.2
	100	70.8	57.2	39.1	18.3	92.2	57.3	22.6	5.3	1.9	98.9
MagFace	5	98.7	95.1	84.6	55.2	77.4	93.6	88.5	82.5	70.6	67.0
	10	97.2	92.3	79.3	47.5	80.9	96.0	92.0	81.5	41.0	88.7
	25	93.2	85.1	67.7	35.7	85.2	98.1	95.4	82.9	41.4	86.8
	50	95.4	86.8	67.4	34.5	85.2	98.8	95.1	79.1	30.8	90.1
	75	93.4	84.6	66.3	35.7	85.3	98.4	93.3	74.8	28.7	89.5
	100	92.4	83.4	65.3	32.9	85.9	98.7	94.6	76.8	25.6	92.0
QMagFace	5	99.8	97.2	86.6	56.4	76.9	97.9	94.1	83.5	60.5	80.8
	10	99.1	94.9	79.7	44.0	82.1	98.3	95.2	89.0	63.1	80.9
	25	98.7	91.7	70.5	36.4	84.6	98.9	94.8	79.1	27.5	91.1
	50	96.4	86.4	63.4	32.6	86.1	98.2	92.4	69.5	20.7	92.3
	75	97.3	90.0	71.0	36.8	84.7	99.4	95.6	77.3	29.2	90.5
	100	97.1	91.1	73.5	39.0	83.7	98.7	93.3	72.5	25.5	91.6

Table 11. Conventional Open-Set Backdoor Performance on WebFace for the digitial sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the digital sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	97.7	94.0	83.7	53.9	78.1	97.8	88.5	61.8	11.4	95.7
	10	95.4	89.0	71.7	39.7	84.7	97.6	85.4	43.6	9.3	96.9
	25	88.3	75.5	54.8	25.8	90.3	96.8	79.7	44.5	9.5	96.8
	50	83.6	69.9	49.0	22.9	91.1	93.2	78.7	41.6	8.2	97.0
	75	83.8	68.4	46.8	21.7	91.3	96.4	78.4	44.9	11.6	96.0
	100	83.9	68.2	47.3	22.3	91.1	92.3	80.3	47.4	11.7	95.9
ArcFace	5	99.1	96.5	87.8	56.9	78.0	89.3	71.4	33.3	1.5	98.6
	10	96.3	89.6	71.9	36.5	86.4	72.2	31.7	3.6	0.0	99.8
	25	91.6	80.2	58.7	27.5	89.2	78.4	32.3	6.4	0.3	99.7
	50	79.7	65.2	44.6	20.3	91.7	54.2	26.0	5.0	0.0	99.8
	75	72.7	57.4	38.3	17.9	92.7	70.6	26.2	7.6	0.9	99.5
	100	68.6	52.3	33.9	15.7	93.5	58.4	27.0	5.8	0.3	99.7
MagFace	5	99.3	97.5	91.9	68.0	70.0	99.5	97.2	87.1	47.2	85.2
	10	97.2	92.4	78.8	45.9	82.3	98.2	87.7	54.2	13.4	95.1
	25	85.6	70.3	49.3	23.8	90.4	89.7	78.0	44.1	11.1	95.4
	50	76.1	58.3	38.6	17.2	92.7	93.4	72.0	38.8	8.4	96.9
	75	70.5	56.4	37.6	19.1	91.6	94.2	64.7	29.2	7.5	97.5
	100	73.9	56.4	38.0	18.4	92.2	90.7	75.5	39.7	8.8	96.6
QMagFace	5	99.9	99.5	94.9	69.9	70.1	99.9	99.8	95.8	66.2	79.4
	10	97.0	91.4	75.0	40.9	83.5	98.9	91.9	64.5	18.2	94.1
	25	84.2	70.5	49.7	23.4	90.4	93.7	75.6	40.1	8.3	97.0
	50	76.0	60.4	40.4	20.4	90.9	95.6	76.8	42.5	7.7	96.9
	75	75.0	58.7	37.9	17.9	92.4	91.2	71.6	39.9	11.2	95.6
	100	73.7	56.1	36.7	16.6	92.8	91.4	74.8	40.3	8.2	97.1

Table 12. Conventional Open-Set Backdoor Performance on WebFace for the hat trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the hat trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	98.3	95.3	84.7	56.5	77.8	99.3	77.6	58.4	7.5	96.9
	10	96.7	90.8	73.9	39.6	84.6	90.2	62.5	65.9	19.0	93.7
	25	89.9	78.2	58.1	28.5	88.4	87.4	46.2	53.2	10.6	96.3
	50	84.5	71.1	50.7	22.6	90.9	84.7	58.0	50.1	6.5	97.2
	75	83.4	69.9	49.1	23.0	90.9	71.4	40.7	14.3	95.0	
	100	85.5	70.4	48.9	22.6	91.1	81.3	53.5	50.6	6.8	97.2
ArcFace	5	99.2	97.1	89.2	60.0	76.1	94.8	77.6	37.4	1.8	98.5
	10	97.3	89.7	72.4	36.7	86.0	90.2	62.5	12.8	0.1	99.5
	25	91.2	80.7	60.3	29.1	88.3	87.4	46.2	12.9	0.6	99.4
	50	79.4	65.2	45.4	20.5	91.8	84.7	58.0	21.1	2.5	98.8
	75	72.2	57.5	37.3	16.8	93.4	71.4	40.7	12.9	0.8	99.4
	100	68.5	51.2	32.7	14.6	93.7	81.3	53.5	22.9	3.6	98.5
MagFace	5	99.2	97.7	91.3	65.1	72.3	99.8	98.2	89.7	59.3	80.7
	10	96.6	90.1	72.2	38.7	85.6	99.3	93.6	73.1	24.9	92.5
	25	82.8	70.1	49.1	22.1	90.8	94.2	81.4	42.9	3.4	98.0
	50	78.8	64.2	43.0	19.2	91.9	97.1	85.5	55.9	12.2	95.8
	75	73.4	57.0	37.3	17.5	92.5	97.3	85.6	56.6	10.7	96.1
	100	70.0	53.3	34.5	16.2	92.8	96.6	81.5	46.6	9.4	96.7
QMagFace	5	99.8	98.8	93.9	73.0	69.4	99.9	99.3	93.3	64.2	79.7
	10	96.5	89.5	72.3	39.7	84.4	99.3	93.1	71.6	20.1	93.7
	25	88.9	73.0	49.7	23.0	90.6	95.0	83.0	51.1	14.5	95.6
	50	79.6	62.8	41.9	19.9	91.3	98.5	86.7	57.8	14.2	95.3
	75	70.6	54.1	34.8	16.7	92.4	95.4	81.7	46.4	7.8	97.0
	100	73.8	56.1	37.0	17.3	92.7	96.3	81.4	54.0	14.8	95.2

Table 13. Conventional Open-Set Backdoor Performance on WebFace for the physical sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the physical sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	98.6	95.2	84.8	55.4	77.2	99.5	97.8	90.5	59.0	81.4
	10	96.7	91.1	75.5	43.8	82.1	97.7	93.7	77.5	30.7	90.4
	25	92.5	83.3	62.7	32.9	85.9	98.9	95.0	78.9	27.5	91.7
	50	92.0	80.7	62.5	33.8	85.8	99.5	96.4	77.6	23.5	92.9
	75	94.4	84.4	65.8	37.3	85.0	99.7	96.9	80.2	27.4	91.6
	100	93.2	82.9	64.9	35.2	85.5	99.7	96.9	79.3	25.9	91.9
ArcFace	5	98.7	95.3	86.0	56.5	77.3	92.3	86.9	81.1	60.8	81.8
	10	98.4	94.9	83.0	51.9	79.3	92.4	85.2	54.3	5.4	97.3
	25	94.9	87.9	70.7	38.4	84.1	94.1	79.5	27.6	1.1	98.9
	50	88.6	76.5	56.0	27.7	88.8	76.4	43.3	8.9	0.4	99.6
	75	80.1	68.1	48.9	23.6	90.0	75.4	35.2	7.9	0.4	99.6
	100	79.0	65.6	45.9	22.2	90.2	64.6	34.8	5.9	0.1	99.8
MagFace	5	99.0	96.4	88.2	60.4	74.4	94.9	92.2	88.4	79.9	50.3
	10	96.6	89.9	72.7	38.7	84.5	95.9	93.6	91.6	87.9	48.3
	25	91.0	79.1	56.9	25.8	89.3	97.9	96.9	95.3	91.0	47.8
	50	94.7	85.6	66.1	32.3	87.3	99.4	98.6	95.4	82.5	55.4
	75	93.9	84.3	64.5	32.8	86.7	99.4	98.8	95.7	85.5	51.4
	100	95.7	87.6	70.6	38.6	84.1	99.7	99.1	95.9	82.4	56.2
QMagFace	5	99.3	97.2	88.5	60.0	75.4	99.3	96.6	86.6	80.7	43.4
	10	99.2	95.7	81.2	45.6	81.8	99.2	97.1	92.4	85.6	55.6
	25	96.6	88.0	67.9	32.2	87.4	99.0	97.7	95.9	89.8	48.1
	50	96.5	86.2	63.9	32.0	87.0	99.7	98.7	94.6	82.5	55.4
	75	95.1	86.0	63.2	31.0	87.6	99.7	98.8	95.8	83.5	54.3
	100	98.6	95.1	78.6	38.2	84.4	99.8	99.4	96.3	81.7	55.3

Table 14. Conventional Open-Set Backdoor Performance on WebFace for the red square trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the red square trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).