Deep transfer learning on the aggregated dataset for face presentation attack detection.

ABDULLAKUTTY, F., ELYAN, E., JOHNSTON, P. and ALI-GOMBE, A.

2022

Cognitive Computation
https://doi.org/10.1007/s12559-022-10037-z

Deep Transfer Learning on the Aggregated Dataset for Face Presentation Attack Detection

Faseela Abdullakutty[1] · Eyad Elyan[1] · Pamela Johnston[1] · Adamu Ali-Gombe[1]

Received: 10 December 2021 / Accepted: 17 June 2022

Abstract

Presentation attacks are becoming a serious threat to one of the most common biometric applications, namely face recognition (FR). In recent years, numerous methods have been presented to detect and identify these attacks using publicly available datasets. However, such datasets are often collected in controlled environments and are focused on one specific type of attack. We hypothesise that a model's accurate performance on one or more public datasets does not necessarily guarantee generalisation across other, unseen face presentation attacks. To verify our hypothesis, in this paper, we present an experimental framework where the generalisation ability of pre-trained deep models is assessed using four popular and commonly used public datasets. Extensive experiments were carried out using various combinations of these datasets. Results show that, in some circumstances, a slight improvement in model performance can be achieved by combining different datasets for training purposes. However, even with a combination of public datasets, models still could not be trained to generalise to unseen attacks. Moreover, models could not necessarily generalise to a learned format of attack over different datasets. The work and results presented in this paper suggest that more diverse datasets are needed to drive this research as well as the need for devising new methods capable of extracting spoof-specific features which are independent of specific datasets.

Keywords Presentation attack detection · Data aggregation · Unseen attacks

Introduction

Face Recognition (FR) has become one of the most popular biometric modalities, having improved significantly over the years. However, FR systems are prone to various attacks, degrading system reliability and security. Presentation Attacks (PA) are ubiquitous among these attacks. Unfortunately, PAs have become a profound threat to FR systems. PAs are the imitation of genuine user faces in the form of a photo, video, or mask. Imposters attempt to circumvent FR systems using such PA variants.

Using these PAs, imposters can either impersonate or obfuscate [1]. Impersonation is the process of attaining access through FR systems, using the replica of genuine facial features. By obscuring a user's identity, obfuscation enables them to pass a security system unnoticed. There are two types of PAs: 2D attacks and 3D attacks [2]. Photo and video attacks are 2D attacks, whereas 3D masks and make-up attacks are 3D attacks. Online portals are attacked using simple attacks, such as photos or videos. During border control scenarios, imposters use more sophisticated attacks such as silicon masks or make-up to fool security systems.

With a wide variety of existing attacks and an unlimited potential for new ones to emerge in the future, PAs are extremely diverse. Photo attacks have different variants, such as warped, printed, eye-cut, and displayed photos [3]. Even within each variant of attacks there will be differences based on domain-dependent features. These domain-dependent features include capturing device, the material used for printing, illumination, resolution, display device, and the physical environment also causes variance. Video attacks have variants based on resolution and display devices. Manufacturers use distinct materials to make masks. Paper masks and wax masks are rigid masks, whereas silicon masks and rubber masks are non-rigid masks [4]. Flexible masks like silicon masks are much harder to detect because of their close similarity to human skin texture and appearance. Photo and video attacks are effortless to reproduce. The seamless access to personal

✉ Faseela Abdullakutty
f.abdullakutty@rgu.ac.uk

[1] Robert Gordon University, Aberdeen, UK

Published online: 07 July 2022

images and videos through social media also helps in replicating 2D attacks with ease. So much diversity among PA techniques presents a significant challenge to PA detection methods. Successful Face Presentation Attack Detection (FPAD) models need to generalise across as many existing PA techniques as possible. In addition, these models should generalise across various domain-dependent features.

FPAD models employed hand-crafted features in early research. These features were classified using traditional machine learning techniques such as Support Vector Machine (SVM) and Random Forest (RF). Local Binary Patterns (LBP) [5–7], Histogram of Oriented Gradients (HOG) descriptors [8, 9], Speeded-Up Robust Features (SURF) [10], and Difference of Gaussian (DoG) [11, 12] are all examples of hand-crafted features. These methods performed well in intra-dataset evaluation with public Face Anti-Spoofing (FAS) datasets. However, hand-crafted features, especially textural features, can be specific to the conditions captured within individual datasets, and the traditional machine learning classifiers further accentuated this.

Convolutional Neural Networks (CNN), with their exceptional inherent feature extraction capability, improved PA detection [13, 14]. CNN-based methods included transfer learning, anomaly detection [15], auxiliary supervision [16, 17], few-shot and zero shot methods [18], and multi-modal methods [19]. Within transfer learning, there are two further categories: domain adaptation [20] and domain generalisation [21].

Although CNN-based models demonstrated impressive intra-dataset performances, they were not able to generalise against unseen attacks. Existing models might have used available public FAS datasets to train the model [22]. These datasets have limited variance in terms of attacks and domain-dependent features. In contrast, attacks are more diversified in real-life scenarios. They differ in attack types and domain features. Hence, models trained on existing datasets may not generalise against such unseen attacks, or even known attack formats in new physical environments. As a result, the reliability of the FR system deteriorates in practical applications. Moreover, emerging novel attacks have become a major threat to the generalisation capability of deployed FPAD models. This has led to further investigation of the generalisation problem in FPAD [23]. One way to tackle the challenge of generalisation in FPAD is to produce a large and comprehensive dataset with many diverse attack variants, simply by aggregating existing datasets. Hence, the impact of data aggregation is investigated in this article, to address generalisation of deep transfer learning models in FPAD context.

This article extensively evaluates the performance and generalisability of models trained on aggregated datasets. Pre-trained VGG-16, ResNet-50, Inception V3, and DenseNet-121 models were trained on NUAA, CASIA, and Replay Attack datasets and their combinations. The prime contribution of this paper is an experimental framework that evaluates how publicly available FAS datasets may be aggregated to enhance inter-dataset performance.

The significant contributions of this article are:

- An experimental framework to assess the generalisation capability of deep transfer learning models.
- An aggregated dataset combining three popular FAS datasets to evaluate PA detection performance of deep transfer learning models.
- An extensive evaluation of the experimental framework using four FAS datasets and their combinations.

The experimental analysis in this paper opens up new research directions to improve generalisation in face presentation attack detection.

The remaining part of the article is as follows: the "Related Works" section analyses existing literature which has used aggregated datasets and transfer learning for face presentation attack detection. The dataset, models, and experiments are described in the "Methods" section. A detailed discussion of the obtained results is presented in the "Results" section. The article is concluded in the "Conclusion" section suggesting future research directions.

Related Works

Presentation attack detection has attained significant improvement over the years, especially with CNN-based models. As described in the "Introduction" section, these models showed reduced generalisation capability against unseen attacks in real-life scenarios when compared with their benchmark statistics. The major cause for this deteriorated performance is the limited variance in training datasets. Hence, unseen attack detection across a wide range of attacks and across different datasets is still considered a challenging problem [23]. Public FAS datasets include only a few attack variants and domain-dependent features, such as illumination, settings, spoofing medium, and recording devices. Many of the existing FPAD models used one of these datasets for training. Hence, the models showed biasing towards the training dataset, exhibiting reduced generalisation against novel attacks.

There have been several studies exploring the concept of data aggregation to address the generalisation problem in FPAD. Costa-Pazo et al. [22] proposed an aggregated dataset to provide more variance in terms of attack types, lighting, recording devices, and resolution. The authors combined ten public datasets to build the GRAD-GPAD (Generalisation Representation over Aggregated Datasets for Generalised Presentation Attack Detection). They also used a uniform protocol to evaluate the colour-based [24] and quality-based [25] models. This framework was further extended in [26] including demographic bias analysis and finer categorisation of PAs based on different factors such as resolution, spoofing medium, and materials. This enhanced aggregated dataset mimicked more realistic scenarios. Thus, the GRAD-GPAD and protocols facilitated evaluation of generalisation capability the-state-of-the-art methods. However, this grand dataset did not include multi-spectral datasets because of data incompatibility.

Saha et al. [27] also addressed domain generalisation using multiple datasets. The authors used four public datasets: Replay Attack [5], CASIA [3], OULU-NPU [28], and MSU-MFSD [29]. Three datasets were included in the training set, whereas the fourth was used for evaluation. The model learned the features from the three training datasets as single domain features. Thus, the model could use more domain-dependent features, leading to better detection performance. Following the concept of dataset aggregation to improve domain generalisation, Nikisins et al. [25] combined three public datasets to illustrate the drawback of binary classification methods in detecting unseen PAs and evaluate their one-class classification model. The authors also established a specific evaluation protocol for the aggregated dataset, combining Replay Attack [5], Replay-Mobile [30], and MSU-MFSD [29]. The train, development, and testing sets were disjoint sets in terms of attacks. The aggregated dataset showed lower half total error rate (HTER) with image quality measure methods when all the PA samples were part of the training set. However, binary classification exhibited poor performance on unseen attack detection. Authors of [31] used CASIA instead of Replay-Mobile to form an aggregated dataset. The authors of [32] and [33] used data aggregation in FPAD. Both of these works combined the real faces from the datasets, keeping the attack faces from each dataset with different domain features dispersed. They adapted this procedure to attain a generalised feature space.

Transfer learning utilises learned knowledge from one task for other similar ones. It assists in mitigating overfitting due to data limitations. Not only that, transfer learning saves computational resources as it avoids training deeper networks from randomised initial parameters. FPAD is a binary classification problem as it identifies if spoofing is present or not, and FPAD datasets are typically visible light spectrum, RGB images. Hence, a deep network that was trained for image classification with datasets like ImageNet [34] can be used to formulate a model to detect PAs. These pre-trained networks were used with fine-tuning either only top layers or a few convolutional layers with top layers.

In [14], Lucena et al. used transfer learning to address the FPAD problem. The authors fine-tuned a VGG-16 model that was pre-trained on ImageNet. Evaluation with a face spoof detection dataset demonstrated improved results over the existing the-state-of-the-art methods. Nagpal and Dubey [13] carried out extensive experiments using different pre-trained models to detect spoofed faces. The authors observed that transfer learning with deep models provided better results than using these networks with random weights or training from the beginning. Yu et al. [35] proposed a face anti-spoofing model using neural architecture search and transfer learning. In [36], the authors used transfer learning and short wave infra-red (SWIR) images for FPAD. A pre-trained face recognition network was used for transfer learning. Authors of [37] adopted a novel method to detect spoofed faces using extracted intrinsic image features and transfer learning. ResNet-50 [38] was used for implementing transfer learning which enhanced spoof detection using the extracted features from the datasets NUAA, CASIA, and Replay Attack. Tu and Fang [39] utilised transfer learning using ResNet-50 and the long short-term memory (LSTM) to address FPAD. Compared to the state-of-the-art methods using feature extraction and shallow networks, these transfer learning-based methods exhibited better detection performance.

George et al. [19] used Light CNN, which is a pre-trained FR model and the concept of domain-specific unit (DSU) to address FPAD. This method utilised a multi-modal dataset with four modalities. The low-level layers were re-trained using the new dataset and re-used the higher level weights. The extracted features from each modality data were concatenated together to form a final feature vector which was then passed to a fully connected layer of size 10 followed by sigmoid layer for classification. In this way, a pre-trained FR model was fine-tuned to adapt to the FPAD task using multi-modal data. Authors of [40] fine-tuned the face recognition CNN model pre-trained on an LWF [41] dataset, similar to the aforementioned [19], to address domain adaptation of PAs in NIR. The initial two convolutional layers and first fully connected layers were made trainable in the fine-tuning. This facilitated the pre-trained model adaptation to the PAD task with NIR images. Even though the models were pre-trained on RGB data, the authors recorded a new NIR dataset with variance in illumination, environmental settings, subject pose, appearance, and attack types. The model was able to detect photo and video attacks better than mask attacks.

The authors used the pre-trained FR model, Light CNN as a backbone/feature extractor to set up patch pooling concept to address FPAD, in [42]. Li et al. [43] proposed another dual mode method using NIR and RGB data to detect spoof. The authors used a light-weight network MobileNet-V3 as the backbone of the model. Each branch of the model was used to extract features from NIR and RGB data separately using this pre-trained model. The selected features were then passed to a softmax layer for classification.

Even though existing literature has explored the concept of combining multiple datasets for training the FPAD model, they rarely included the NUAA imposter dataset. Various handcrafted feature methods and deep learning methods were evaluated using NUAA. Either official or customised partitions were used to evaluate these methods [44]. This article used transfer learning and the concept of data aggregation to address the generalisation in FPAD. The experiments used a combined training set of official training partitions from NUAA, CASIA, and Replay Attack. These three datasets have distinct 2D attack variants and domain-dependent features.

Methods

The experimental framework in this article used transfer learning with binary classification to perform FPAD. Pre-trained deep networks, VGG-16 [45], ResNet-50 [38], Inception V3 [46], and DenseNet-121 [47] were used for transfer learning.

Three widely known public datasets, NUAA [12], Replay Attack [5], and CASIA [3], were considered for these experiments and forming aggregated datasets. These three datasets and their combinations were used for training. All three datasets followed their official train/test split. Real face images from the three datasets combined to form a real face class in an aggregated dataset. Similarly, an attack class also was formed using attack images from these datasets. The combined train set provided different attack variants. For cross-dataset evaluation on this aggregated training set, SiW [48] test set was also used.

Aggregated Dataset

This experimental framework focuses on examining the impact of data aggregation on the generalisation of FPAD. To accomplish this task, an aggregate train set was constructed with NUAA, CASIA, and Replay Attack datasets' train partitions. NUAA consists of print attacks. Replay Attack has video attacks. CASIA includes both photo and video attacks. CASIA has warped, print, eye-cut photo attack variants. Thus, the resulting aggregated train set has both video and photo attacks with variance in attack types and domain-dependent features.

The number of images in each class corresponding to three datasets and the combined dataset is shown in Table 1. In Fig. 1, the distributions of real and fake classes in individual and aggregated dataset are presented. NUAA has almost equal number of real and fake class images in train set. However, CASIA and Replay Attack have more fake face images than real ones in train set. The aggregated train set includes 3959 real and 8769 fake face images.

Table 1 Number of real and fake images in three datasets and aggregated dataset

Dataset	Train		Test	
	Real	Fake	Real	Fake
NUAA	1743	1748	3362	5761
Replay Attack	1689	5261	1928	5645
CASIA	527	1760	824	2471
Aggregated dataset	3959	8769	6114	13877

Given the unique characteristics of each dataset, such as lighting, the spoofing medium, the environment, and the recording device, the combination of these datasets produces greater variance in both real and fake classes. As a result, the model can learn a wider range of features to distinguish between real and attack classes. Furthermore, it avoids overfitting due to subtle biases within a single dataset. The individual training sets from each dataset were combined to form a training set, as indicated in Fig. 1. The aggregated test set was also constructed in this manner. Each dataset was divided following the official train/test protocol. Consequently, no mixing up of train and test set distributions occurred in the aggregated dataset. By keeping the distributions consistent with the official protocol, even in the aggregated dataset, we maintain the challenges of domain generalisation inherent to individual datasets [44].

CNN Models

An FPAD determines the authenticity of detected faces. In essence, it involves binary image classification. With deep CNN models, FPAD has also achieved significant improvement, similar to any other computer vision task [13, 14, 49]. It must be noted, however, that deep neural networks require a substantial quantity of data to achieve

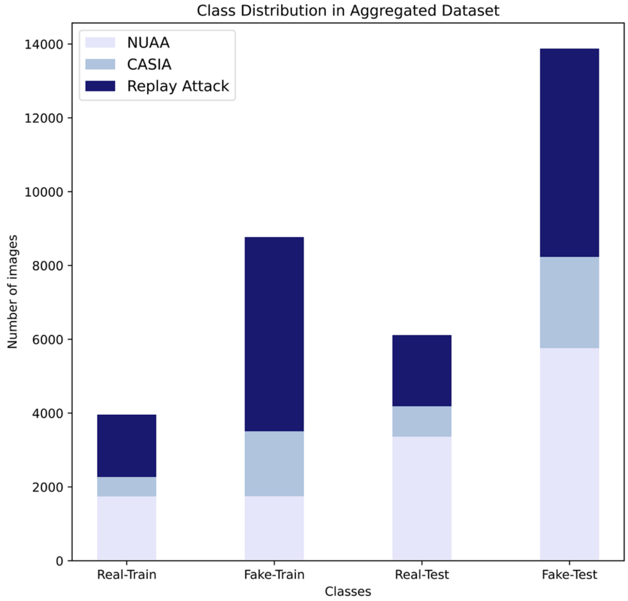


Fig. 1 Class distribution in aggregated dataset

desired performance. In order to solve such problems, transfer learning has become increasingly popular. By freezing some layers of the network and retraining others on a new dataset from the new domain, transfer learning re-purposes an already learned network to perform a similar task. In this manner, a task may be accomplished with less training data, less time, and with higher accuracy. The majority of the FAS datasets are restricted in size. As a result, transfer learning was used to overcome this limitation.

The experimental framework in this article utilises transfer learning to evaluate the aggregated dataset performance. For this purpose, we used pre-trained deep neural networks with architecture VGG-16 [14], ResNet-50 [37, 39], Inception V3 [13], and DenseNet-121 [16]. These models were popularly used and experimentally verified for FPAD in existing literature [49]. The networks used in the experiments were all pre-trained using ImageNet [34]. Pre-trained models were loaded without output layers, freezing the top layers. The models used "ImageNet" weights. Top layers were fine-tuned using FAS datasets to perform PA detection.

Experiments

Extensive experiments were carried out to evaluate the PA detection with the aggregated dataset. The considered models were trained using three individual datasets, NUAA, CASIA, and Replay Attack and their various combinations. To assess the generalisation capability, cross-dataset evaluation was also performed using a fourth dataset (SiW) test split. Thus, the impact of data aggregation was evaluated using this experimental framework and four datasets. The FAS datasets, transfer learning models, hyper-parameters, and experimental settings are described in the "Datasets" section and "Experimental Settings" section.

Datasets

The experiments used three FAS datasets. They were NUAA, CASIA-FASD, and Replay Attack. In existing literature, both traditional hand-crafted feature extraction methods and recent deep learning methods in FPAD have used these three datasets for evaluation. The different attack variants, test protocols, and lighting conditions also assist in creating more variance within the aggregated dataset. CASIA and Replay Attack datasets, as distributed, consist of videos. Frames were extracted with a rate of 2 fps and face detection was carried out on these frames. NUAA was accessed as face detected images, which are provided as part of the official dataset. These face detected images were resized to 224 × 224 pixels. Official test/train partitions were used for each dataset. The experiments also used an SiW test set to perform cross-dataset evaluation on a combined train set consisted of NUAA, Replay Attack, and CASIA train sets. The facial images were extracted at a frame rate of 1 fps from each video to form this dataset. The SiW train set was unused. Table 2 shows the number of train and test images in each dataset, which were used in the experiments.

The NUAA imposter database contains authentic images as well as photo attack and covers samples of 15 individual subjects. In contrast to the training set, the official test set is considerably larger. The training set contains 3491 images, while the test set contains 9123 images. These images were extracted from videos recorded at three different sessions under different lighting conditions. However, the already extracted images after face detection are available to the public. In NUAA's train partition, both classes have nearly the same number of images, whereas in the test set the attack images are much more numerous than the real face images. In terms of attack variants, it consists only of photo attacks. Despite these facts, NUAA remains popular among FPAD researchers [49–51].

CASIA-FASD has print attacks, warped photo attacks, cut photo attacks, and video attacks. Fifty subjects were represented with fake and real faces. There are three real face videos and nine fake face videos for each subject. The train set features 20 people. There are genuine and fake videos of 30 individuals in the test set. The train and test sets are disjoint in terms of subjects. There are three real face videos and 9 attack videos corresponding to each subject. Thus, the train set has 60 real face and 180 attack videos in total. The test set includes 90 real face and 270 attack videos. Like NUAA, CASIA lacks ethnically diverse subjects. In addition, CASIA includes seven test cases, including three attack types, three image quality levels, and the entire dataset. In this experiment, the entire dataset is used with the given partition based on the dataset protocol. As this dataset has more attack variants, including video attacks, it is widely used for evaluating FPAD models [52–54].

Replay Attack was created by using 50 identities. There were respectively 15 subjects for training, 15 subjects for development, and 20 subjects for testing. While recording the Replay Attack dataset, printed images, mobile displays, and tablets were utilised. The three mediums were either fixed to a support or held by the operator during the recording process. Two types of recording environments were used to capture the videos, controlled and adverse. The controlled setting had a uniform background

Table 2 Number of train and test images in each dataset

Dataset	Train	Test
NUAA	3491	9123
Replay Attack	6950	7573
CASIA	2287	3295
SiW	40790	34779

and illumination using incandescent lamps, whereas the adverse setting had a non-uniform background and daylight illumination. There are various PA types in this dataset. Hence, it is popular among the FPAD researchers [55, 56]. Both train and development sets have 60 real face videos and 300 attack videos. The test set consists of 80 real face videos and 400 attack videos.

The Spoof in the Wild (SiW) [48] dataset consists of 165 subjects from a more diversified ethnicity than the other datasets. There are 8 real face and 20 attack videos corresponding to each subject. Thus, the dataset has 4620 videos. The dataset was made using 6 spoofing mediums. Four different sessions were used varying factors such as poses, illuminations, expressions (PIE), and distance-to-camera. Videos were pre-processed by first using the frame rate to extract one image per second. Then, the face area was extracted using the annotations provided. To increase diversity of facial images, the face area was cropped to accommodate some background information. This was achieved by multiplying each bounding box with a random scaling factor between 1.1 and 1.4. Finally, images were resized to 224 × 224.

Experimental Settings

The experiments included intra-dataset and cross-dataset evaluations using individual datasets and their different combinations. To carry out cross-dataset evaluation on the aggregated train set, the SiW test set was used. Thus, each model was evaluated using the dataset combinations as in Table 3. Models were trained for 10 epochs with a batch size of 32. These parameters were the same for all classification models. A validation split of 20% of the train set was used while training the model. To compare the performance in binary classification ROC curve, accuracy, half total error rate (HTER), precision, recall, F1 score, false positive (FP), and false negative (FN) are used. The HTER is the average of false acceptance rate (FAR) and false rejection rate (FRR).

As presented in Table 1 and Fig. 1, two classes from three datasets were combined, and this aggregated train set was used to train the four models. For binary classification, there were real and fake classes irrespective of the actual dataset. The output layers in the base pre-trained model were replaced with one dense layer with a size of 1000 and sigmoid activation function followed by a softmax classification layer, size 2. Binary cross-entropy was used as a loss function. The Adam optimiser [57] was used with all four models. The learning rate for VGG-16, ResNet-50, and DenseNet-121 was 10^{-5}. For Inception V3, the learning rate was 10^{-6}.

Table 3 Test set v/s train set combinations used in the evaluation

No.	Train set	Test set
1	NUAA	NUAA
2		Replay Attack
3		CASIA
4	Replay Attack	NUAA
5		Replay Attack
6		CASIA
7	CASIA	NUAA
8		Replay Attack
9		CASIA
10	NUAA+CASIA	Replay Attack
11	NUAA+Replay Attack	CASIA
12	CASIA+Replay Attack	NUAA
13	NUAA+CASIA+Replay Attack	Replay Attack
14	NUAA+CASIA+Replay Attack	CASIA
15	NUAA+CASIA+Replay Attack	NUAA
16	NUAA+CASIA+Replay Attack	SiW

Results

Extensive experiments and analysis were performed to investigate the influence of dataset aggregation in face presentation attack detection. Considered pre-trained models in the experiments were trained with three public FAS datasets and their combinations as in Table 3. Both intra-dataset and cross-dataset evaluations were carried out to compare the model performance in FPAD. Intra-dataset evaluation results using individual and the aggregated datasets are presented in Table 4. Receiver operating characteristic curve (ROC) comparison of each model with all three datasets in intra- and cross-dataset evaluation scenarios is presented in Figs. 2 and 3. CASIA (93.35%) and Replay Attack (95.89%) showed the best intra-dataset performance with DenseNet-121 in intra-dataset evaluation. In contrast, NUAA had the highest performance (82.61%) with ResNet-50 in the intra-dataset evaluation.

The experimental framework also trained models with aggregated datasets. These models were then tested with individual test sets from CASIA, Replay Attack, NUAA, and with an aggregated test. In the aggregated dataset evaluations, NUAA datasets exhibited the lowest accuracy with more than 40% HTER (Table 4). False Positive Rate (FPR) increased in the aggregated dataset evaluation on NUAA test sets. This increase in FPR caused the lower accuracy and higher HTER for NUAA with all four model architectures. On the other hand, CASIA and Replay Attack showed a decrease in FPR in the aggregated dataset evaluation. ResNet-50 with Replay Attack was an exception, where FPR increased from 16.65 to 18.52%. As the FPR increased, it lowered the accuracy

Table 4 Comparison of intra-dataset and the aggregated dataset evaluations. The highlighted values show the best performance, when tested with each dataset in intra and aggregated dataset evaluation

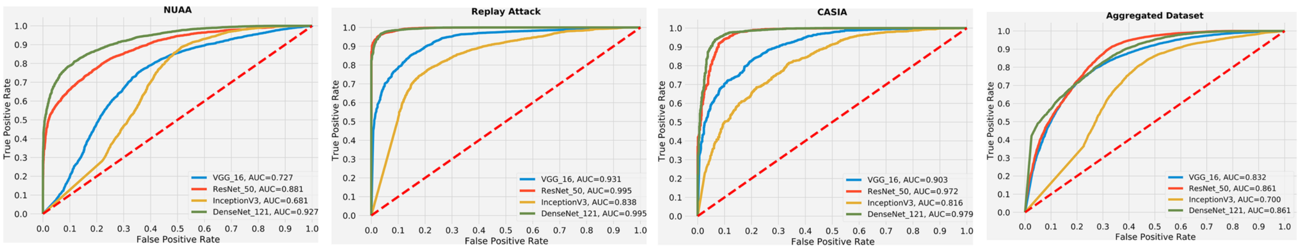
Train	Test	VGG-16		ResNet-50		Inception V3		DenseNet-121	
		ACC (%)	HTER (%)	ACC (%)	HTER (%)	ACC (%)	HTER (%)	ACC (%)	HTER (%)
NUAA	NUAA	73.19	28.41	**82.61**	**19.08**	67.48	37.00	80.79	17.40
Replay Attack	Replay Attack	86.61	20.36	95.34	8.61	81.44	29.43	**95.89**	**6.76**
CASIA	CASIA	86.97	22.14	92.92	13.61	81.98	27.16	**93.35**	**12.85**
Aggregated data	NUAA	**72.83**	**31.74**	71.18	38.72	67.97	40.40	64.61	42.71
Aggregated data	Replay Attack	86.84	18.69	93.72	10.30	79.50	27.64	**97.32**	**4.37**
Aggregated data	CASIA	86.45	20.16	92.02	11.66	81.69	28.83	**93.42**	**12.21**
Aggregated data	Aggregated data	79.981	25.770	83.177	26.574	74.584	34.100	81.447	25.528

slightly. With DenseNet-121, both FPR and False Negative Rate (FNR) decreased for Replay Attack in the aggregated dataset evaluation. These decreased FPR and FNR facilitated performance improvement in this specific evaluation. On the other hand, for CASIA, despite the decreased FPR, FNR doubled in the same evaluation scenario, resulting only a slight improvement in accuracy (93.35% to 93.42%).

Cross-dataset evaluation results are shown in Table 5. Evaluation was carried out using individual datasets and their combinations for training (Table 3). To evaluate the performance of the aggregated dataset, an SiW test set was also used. The corresponding ROC is shown in Fig. 3d. It is evident from the plot that the cross-dataset performance of the aggregated dataset is significantly low compared to both intra-dataset performance and testing with other individual test. It would seem that SiW is different enough from NUAA, CASIA, and Replay Attack that even an aggregate training set will not enhance generalisation to any great extent. The FPR in detection is more than 50% in most of the testing scenarios regardless of the datasets used, which caused higher HTER.

When trained with aggregated train set and tested with aggregated test set, the models had FPR more than 40%. This FPR value was much greater than the FPR value of testing scenarios with Replay Attack and CASIA test sets. Adding the NUAA dataset while forming the aggregated dataset adversely effected the detection performance. This intra-dataset performance using aggregated datasets can be clearly demonstrated using the corresponding ROC, as in Fig. 2d. Unlike the intra-dataset evaluation on individual datasets, aggregated dataset performance diminished, even with DenseNet-121 (Table 4). For ResNet-50 and VGG-16, this aggregated data intra-dataset performance is near to NUAA intra-dataset performance. However, compared to the other two datasets, the overall intra-dataset performance using the aggregated dataset is low.

It is evident from Figs. 3 and 2 that DenseNet-121 was the best model in both intra- and the aggregated dataset evaluations for CASIA and Replay Attack. However, it was ResNet-50 for NUAA rather than DenseNet-121. NUAA performed the best in the aggregated dataset evaluation with the VGG-16 model. It is evident from the plots that performance on NUAA is not as good as the other two datasets in both the evaluation scenarios. The cross-dataset performance evaluation using the SiW dataset on the aggregated train set was even worse compared to testing the same model with other individual test sets. All the models exhibited very low detection performance in this specific evaluation. This indicates that data aggregation alone does not help generalisation against various attacks.



(a) NUAA **(b)** Replay Attack **(c)** CASIA **(d)** Aggregated dataset

Fig. 2 Intra-dataset evaluation ROC corresponding to individual and aggregated dataset

Table 5 Cross-dataset evaluation results. The highlighted values represent the highest cross-dataset evaluation performance, of the model trained with the aggregated dataset

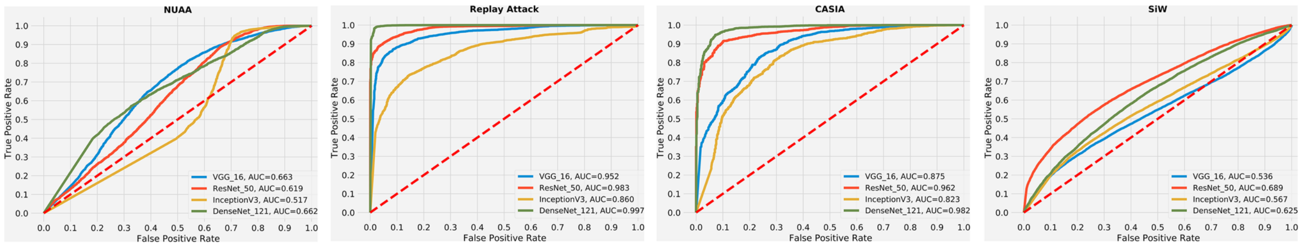
Train	Test	VGG-16		ResNet-50		Inception V3		DenseNet-121	
		ACC (%)	HTER (%)	ACC (%)	HTER (%)	ACC (%)	HTER (%)	ACC (%)	HTER (%)
NUAA	Replay Attack	49.95	50.46	56.68	49.81	32.13	45.54	54.38	41.10
	CASIA	53.78	41.45	54.48	44.35	54.78	33.43	71.08	30.12
CASIA	NUAA	68.39	34.07	59.05	52.94	69.30	37.53	59.37	50.38
	Replay Attack	69.51	48.66	67.42	52.14	50.44	47.47	73.17	50.32
Replay Attack	NUAA	57.71	52.24	59.26	53.08	63.38	49.68	61.93	50.96
	CASIA	30.05	49.03	58.03	57.67	74.48	50.14	68.47	53.18
NUAA+CASIA	Replay Attack	65.81	51.62	68.44	49.07	46.36	41.29	68.16	50.37
NUAA+Replay Attack	CASIA	66.04	40.64	46.56	51.57	60.36	43.82	68.13	45.23
CASIA+Replay Attack	NUAA	61.94	49.30	61.76	51.07	64.69	46.37	53.44	55.76
Aggregated data	SiW	50.49	46.06	64.50	48.78	57.96	46.17	**62.87**	**38.79**

Discussion

The aggregated dataset and cross-dataset evaluation results show that detection rates were reduced when tested with CASIA, Replay Attack, NUAA, and SiW test sets when compared with models trained and tested on a single dataset. It clearly indicates that even though dataset attack variance and size improves with the aggregated dataset, it generally does not improve the detection performance on any component dataset. In fact, combining these datasets led to an increased FPR. Training with these combined datasets restricts the models from identifying real faces correctly. FPAD relies more on spoofing patterns and image quality features. As NUAA was recorded using a webcam, the image quality is lower compared to other datasets in the experiments. Similarly, CASIA also has images of three different qualities, including lower quality images. This quality variation in images influences the high-frequency feature extraction while training the model. Regarding transfer learning, the deep networks used in the experiment were pre-trained for image classification tasks. They extract deep, global features. However, FPAD may require shallow, local features to detect spoofing. These pre-trained image classification models might have failed to learn spoof-specific features to achieve better detection performance, instead relying on some dataset-specific features.

In all the evaluation scenarios, false positives were more significant than false negatives. This shows that even though attacks were detected, the models failed in identifying the genuine images, particularly those in NUAA. This influences the overall performance of these models. CASIA and Replay Attack facial images were extracted from raw videos using the same pre-processing methods. NUAA is available to the public as pre-processed face detected images. These images were resized for experiments. This disparity between the NUAA dataset and other two datasets images may have influenced the classification performance.

The classification was carried out using four deep networks with different architectures. However, except DenseNet-121, other three models exhibited the same performance trend in the aggregated dataset evaluation scenario: the models trained and tested on the same datasets performed better than models trained on an aggregate dataset. Even DenseNet-121 followed the same trend with the NUAA dataset. This implies that combining source domain solely cannot improve the detection performance. The cross-dataset evaluation results presented in Table 5 support this.



(a) NUAA **(b)** Replay Attack **(c)** CASIA **(d)** SiW

Fig. 3 ROC of models trained on aggregated train set and tested on NUAA, Replay Attack, CASIA, and SiW test sets

With the combination of more datasets, handcrafted features were evaluated for generalisation capabilities [22] within the context of FPAD. Based on the results of this research, it was found that state-of-the-art methods with impressive intra-dataset performance are less generalisable in cross-dataset evaluation when used with a combination of heterogeneous sources. A variety of factors influence their performance, including their capture devices, display conditions, and image quality. In contrast to this evaluation, the experiments in this article used binary classification using four pre-trained deep neural networks to detect PAs. It was evident from the analysis of experimental results that even deep learning frameworks were not capable of generalising to different distributions.

Conclusion

Face presentation attack detection was carried out using different publicly available datasets and their combinations. Binary classification using transfer learning was utilised to detect attacks. For training, individual datasets and their combinations were used. An aggregated training set using the official training partitions of NUAA, CASIA, and Replay Attack was also formed to investigate the effect of data aggregation in FPAD generalisation. On the transfer learning models, both intra-dataset and cross-dataset evaluations were carried out. The NUAA dataset exhibited lower performance compared to other two datasets in intra-dataset, aggregated dataset and cross-dataset evaluations. The detection performance reduced when the models were trained with the aggregated training set and tested with test partitions from individual datasets. This shows that combining various source domains only is not sufficient to attain domain generalisation against unseen attacks. Our future work will examine how a model can be customised in such a way that it extracts the spoof-specific and domain invariant features to attain generalisation in this scenario. The results of this extensive experimental analysis show that there is much yet to be learned from further research in areas such as data imbalance, fine-tuning for domain adaptation, compression, and image quality in FPAD context.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors. The research used publicly available datasets with suitable reference.

Conflict of Interest The authors declare no competing interests.

References

1. Singh M, Singh R, Ross A. A comprehensive overview of biometric fusion. Information Fusion. 2019;52:187–205.
2. Ramachandra R, Busch C. Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Computing Surveys (CSUR). 2017;50(1):1–37.
3. Zhang Z, Yan J, Liu S, Lei Z, Yi D, Li SZ. A face antispoofing database with diverse attacks. In: 2012 5th IAPR International Conference On Biometrics (ICB). IEEE; 2012. p. 26–31.
4. Jia S, Guo G, Xu Z. A survey on 3D mask presentation attack detection and countermeasures. Pattern Recogn. 2020;98:107032.
5. Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG). IEEE; 2012. p. 1–7.
6. Määttä J, Hadid A, Pietikäinen M. Face spoofing detection from single images using micro-texture analysis. In: 2011 International Joint Conference On Biometrics (IJCB). IEEE; 2011. p. 1–7.
7. Patel K, Han H, Jain AK. Secure face unlock: spoof detection on smartphones. IEEE transactions on information forensics and security. 2016;11(10):2268–83.
8. Komulainen J, Hadid A, Pietikäinen M. Context based face anti-spoofing. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE; 2013. p. 1–8.
9. Yang J, Lei Z, Liao S, Li SZ. Face liveness detection with component dependent descriptor. In: 2013 International Conference on Biometrics (ICB). IEEE; 2013. p. 1–6.
10. Boulkenafet Z, Komulainen J, Hadid A. Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Process Lett. 2016;24(2):141–5.
11. Peixoto B, Michelassi C, Rocha A. Face liveness detection under bad illumination conditions. In: 2011 18th IEEE International Conference on Image Processing. IEEE; 2011. p. 3557–3560.
12. Tan X, Li Y, Liu J, Jiang L. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: European Conference on Computer Vision. Springer; 2010. p. 504–517.
13. Nagpal C, Dubey SR. A performance evaluation of convolutional neural networks for face anti spoofing. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE; 2019. p. 1–8.
14. Lucena O, Junior A, Moia V, Souza R, Valle E, Lotufo R. Transfer learning using convolutional neural networks for face anti-spoofing. In: International Conference Image Analysis and Recognition. Springer; 2017. p. 27–34.
15. Baweja Y, Oza P, Perera P, Patel VM. Anomaly detection-based unknown face presentation attack detection. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). IEEE; 2020. p. 1–9.
16. Yu Z, Li X, Niu X, Shi J, Zhao G. Face anti-spoofing with human material perception. In: European Conference on Computer Vision. Springer; 2020. p. 557–575.
17. Yu Z, Zhao C, Wang Z, Qin Y, Su Z, Li X, et al. Searching central difference convolutional networks for face anti-spoofing. In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. p. 5295–5305.

18. Liu Y, Stehouwer J, Jourabloo A, Liu X. Deep tree learning for zero-shot face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 4680–4689.
19. George A, Mostaani Z, Geissenbuhler D, Nikisins O, Anjos A, Marcel S. Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans Inf Forensics Secur. 2019;15:42–55.
20. Mohammadi A, Bhattacharjee S, Marcel S. Domain adaptation for generalization of face presentation attack detection in mobile settengs with minimal information. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2020. p. 1001–1005.
21. Zhang KY, Yao T, Zhang J, Tai Y, Ding S, Li J, etal. Face anti-spoofing via disentangled representation learning. In: European Conference on Computer Vision. Springer; 2020. p. 641–657.
22. Costa-Pazo A, Jiménez-Cabello D, Vázquez-Fernández E, Alba-Castro JL, López-Sastre RJ. Generalized presentation attack detection: a face anti-spoofing evaluation proposal. In: 2019 International Conference on Biometrics (ICB). IEEE; 2019. p. 1–8.
23. Abdullakutty F, Elyan E, Johnston P. A review of state-of-the-art in Face Presentation Attack Detection: from early development to advanced deep learning and multi-modal fusion methods. Information Fusion. 2021.
24. Boulkenafet Z, Komulainen J, Hadid A. Face spoofing detection using colour texture analysis. IEEE Trans Inf Forensics Secur. 2016;11(8):1818–30.
25. Nikisins O, Mohammadi A, Anjos A, Marcel S. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In: 2018 International Conference on Biometrics (ICB). IEEE; 2018. p. 75–81.
26. Costa-Pazo A, Pérez-Cabo D, Jiménez-Cabello D, Alba-Castro JL, Vazquez-Fernandez E. Face presentation attack detection. A comprehensive evaluation of the generalisation problem. IET Biometrics. 2021;10(4):408–429.
27. Saha S, Xu W, Kanakis M, Georgoulis S, Chen Y, Paudel DP, et al. Domain agnostic feature learning for image and video based face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020. p. 802–803.
28. Boulkenafet Z, Komulainen J, Li L, Feng X, Hadid A. OULU-NPU: a mobile face presentation attack database with real-world variations. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE; 2017. p. 612–618.
29. Wen D, Han H, Jain AK. Face spoof detection with image distortion analysis. IEEE Trans Inf Forensics Secur. 2015;10(4):746–61.
30. Costa-Pazo A, Bhattacharjee S, Vazquez-Fernandez E, Marcel S. The replay-mobile face presentation-attack database. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE; 2016. p. 1–7.
31. Xiong F, AbdAlmageed W. Unknown presentation attack detection with face RGB images. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE; 2018. p. 1–9.
32. Liu M, Mu J, Yu Z, Ruan K, Shu B, Yang J. Adversarial learning and decomposition-based domain generalization for face anti-spoofing. Pattern Recogn Lett. 2022;155:171–7.
33. Shi L, Zhang J, Liang C, Shan S. Unknown aware feature learning for face forgery detection. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE; 2021. p. 1–5.
34. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference On Computer Vision and Pattern Recognition. IEEE; 2009. p. 248–255.
35. Yu Z, Wan J, Qin Y, Li X, Li SZ, Zhao G. NAS-FAS: static-dynamic central difference network search for face anti-spoofing. arXiv preprint arXiv:2011.02062. 2020.
36. Heusch G, George A, Geissbühler D, Mostaani Z, Marcel S. Deep models and shortwave infrared information to detect face presentation attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2020;2(4):399–409.
37. Bresan R, Beluzo C, Carvalho T. Exposing presentation attacks by a combination of multi-intrinsic image properties, convolutional networks and transfer learning. In: International Conference on Advanced Concepts for Intelligent Vision Systems. Springer; 2020. p. 153–165.
38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition. 2016. p. 770–778.
39. Tu X, Fang Y. Ultra-deep neural network for face anti-spoofing. In: International Conference on Neural Information Processing. Springer; 2017. p. 686–695.
40. Kotwal K, Bhattacharjee S, Abbet P, Mostaani Z, Wei H, Wenkang X, et al. Domain-specific adaptation of CNN for detecting face presentation attacks in NIR. IEEE Transactions on Biometrics: Behavior, and Identity Science. 2022.
41. Huang GB, Mattar M, Berg T, Learned-Miller E. 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.
42. Kotwal K, Marcel S. CNN patch pooling for detecting 3D mask presentation attacks in NIR. In: 2020 IEEE International Conference on Image Processing (ICIP). IEEE; 2020. p. 1336–1340.
43. Li L, Gao Z, Huang L, Zhang H, Lin M. A dual-modal face anti-spoofing method via light-weight networks. In: 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID). IEEE; 2019. p. 70–74.
44. Abdullakutty F, Elyan E, Johnston P. Face spoof detection: an experimental framework. In: International Conference on Engineering Applications of Neural Networks. Springer; 2021. p. 293–304.
45. Zhang X, Zou J, He K, Sun J. Accelerating very deep convolutional networks for classification and detection. IEEE Trans Pattern Anal Mach Intell. 2015;38(10):1943–55.
46. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 2818–2826.
47. Huang G, Liu Z, Van DerMaaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 4700–4708.
48. Liu* Y, Jourabloo* A, Liu X. Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: Proceeding of IEEE Computer Vision and Pattern Recognition. Salt Lake City, UT; 2018.
49. Satapathy A, Livingston LJ. A lite convolutional neural network built on permuted Xceptio-inception and Xceptio-reduction modules for texture based facial liveness recognition. Multimed Tools Appl. 2021;80(7):10441–72.
50. Koppikar U, Sujatha C, Patil P, Hiremath P. Face liveness detection to overcome spoofing attacks in face recognition system. In: Innovations in Computational Intelligence and Computer Vision. Springer; 2021. p. 351–360.

51. Hadiprakoso RB, Setiawan H, et al. Face anti-spoofing using CNN classifier & face liveness detection. In: 2020 3rd International Conference on Information and Communications Technology (ICOIACT). IEEE; 2020. p. 143–147.
52. Wang Y, Song X, Xu T, Feng Z, Wu XJ. From RGB to depth: domain transfer network for face anti-spoofing. IEEE Trans Inf Forensics Secur. 2021.
53. Sharma D, Selwal A. A face anti-spoofing approach based on generic sequential model using scale invariant features. In: 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE; 2021. p. 1–6.
54. Chen B, Yang W, Wang S. Generalized face antispoofing by learning to fuse features from high- and low-frequency domains. IEEE MultiMedia. 2021;28(1):56–64.
55. Zhang Z, Jiang C, Zhong X, Song C, Zhang Y. Two-stream convolutional networks for multi-frame face anti-spoofing. arXiv preprint arXiv:2108.04032. 2021.
56. Huang X, Xia J, Shen L. One-class face anti-spoofing based on attention auto-encoder. In: Chinese Conference on Biometric Recognition. Springer; 2021. p. 365–373.
57. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.