Unmasking the imposters: task-specific feature learning for face presentation attack detection.

ABDULLAKUTTY, F., ELYAN, E. and JOHNSTON, P.

2023

Unmasking the Imposters: Task-specific feature learning for face presentation attack detection

Faseela Abdullakutty
School of Computing
Robert Gordon University
Aberdeen, UK
f.abdullakutty@rgu.ac.uk

Eyad Elyan
School of Computing
Robert Gordon University
Aberdeen, UK
e.elyan@rgu.ac.uk

Pamela Johnston
School of Computing
Robert Gordon University
Aberdeen, UK
p.johnston2@rgu.ac.uk

***Abstract*—Presentation attacks pose a threat to the reliability of face recognition systems. A photograph, a video, or a mask representing an authorised user can be used to circumvent the face recognition system. Recent research has demonstrated high accuracy in intra-dataset evaluations using existing face presentation attack detection models. Nonetheless, these models did not achieve similar performance when evaluated across datasets due to limited generalisation. Consequently, this article presents task-specific feature learning using deep pre-trained models. Model performance was evaluated using three public datasets: the SiW dataset was used for intra-dataset evaluation, while CASIA and Replay Attack were used for cross-dataset evaluation. Custom task-specific feature learning, compared to deep and hybrid models, demonstrated improved cross-dataset performance and exhibited more generalisability. The results suggest future direction for further research toward improving the model's generalisation using custom task-specific feature learning.**

***Index Terms*—Face presentation attack detection, deep learning, generalisation**

I. Introduction

Face Recognition (FR) has become more prevalent as means of personal authentication. Despite significant technological advances, FR systems are still vulnerable to a variety of attacks. The most common attacks involve photos, videos, or masks of authorized users, called Presentation Attacks (PA). With the availability of personal images on the Internet and advances in printing technology, PAs have become more varied and accessible. Face Presentation Attack Detection (FPAD) is extensively studied since these attacks impact the reliability of FR systems. In addition, newly developed PA variants act as unseen attacks against existing FPAD models. In the presence of unseen attacks, detection performance deteriorates, indicating that the models are not sufficiently generalised to capture all types of attacks. Therefore, generalisation has become increasingly significant in the FPAD context [1].

The FPAD model detects PA based on differences in features between fake and genuine facial images. In order to accomplish this, earlier FPAD models used handcrafted features [2] related to texture, image quality, motion, and frequency. The extracted features were classified with SVM, RF or K-NN classifiers. These hand-crafted features are domain-specific features [3]. Hence, hand-crafted feature methods had limited generalisation as they use only domain-specific features rather than task-specific features, especially in the RGB domain.

The automatic feature extraction capability of deep learning models further enhanced FPAD performance. In a deep model, lower layers provide domain-specific features such as edges, and corners. However, the higher layers learn task-specific features. The FPAD task is to differentiate between real and fake facial images of the same user. But image classification categorises different objects in the given images. Thus, task-specific features of FPAD are different from that of image classification. Hence, learning task-specific features is more important in improving generalisation in FPAD.

As a deep learning technique, transfer learning has been exploited in a number of ways to address FPAD by learning either domain-specific features or task-specific features. For transfer learning, existing FPAD models have used pre-trained image classification or face recognition models. Since task-specific features are provided by higher layers, image classification models were used after modifying the top fully connected layers and fine-tuning them to detect PA [4]. Domain-specific features were learned [5] by fine-tuning a few lower convolutional layers in a pre-trained face recognition model using multi-spectral data. Nonetheless, the majority of Face Anti-Spoofing (FAS) datasets are in the RGB domain. So, it may be more effective to use a model that can extract task-specific features from RGB datasets rather than using multi-modal data. The other research [6] has shown that fusion models using deep pre-trained models and hand-crafted methods improved PA detection in intra-dataset evaluations. Thus transfer learning has been explored extensively in face anti-spoofing.

This article presents a transfer learning model, to learn task-specific features to improve generalisation. The higher convolutional layers of deep pre-trained models were fine-tuned along with the fully connected layers using a public FAS dataset SiW. This fine-tuned model was used to extract features, which were used to form fusion models. Fusion models were formed using the deep features from fine-tuned models and combining the deep features with hand-crafted features. The experiments used the public FAS datasets, CASIA, and Replay Attack, for cross-dataset validation. The main contributions of this article are:

- A task-specific feature learning method using deep pre-trained models for Face Presentation Attack Detection (FPAD).
- Fusion models combining deep fine-tuned features and hand-crafted features for FPAD.
- Intra-dataset and cross-dataset evaluations to assess the detection performance and generalisability of the models.

II. Related Work

Face recognition remains vulnerable to PA despite recent technological advancements. In order to create PAs, new technologies are being utilized, including manufacturing and printing materials, lighting, and cameras with improved resolution, to name a few. There is a serious problem associated with the generalisability of the existing FAS models arising from these emerging PAs [1]. This has driven the biometric research community to investigate a variety of techniques to improve the generalisation of FPAD. Thus, FPAD can be divided into three types: hand-crafted feature methods, deep learning-based methods and hybrid methods. Many existing FPAD models have challenges related to generalisation, particularly when tested against unseen attacks.

FPAD uses feature descriptors related to texture, colour, image quality and frequency. Extracted features are passed to standard classifiers such as SVM (Support Vector Machine) and RF (Random Forest) for PA detection. Local Binary Pattern (LBP) and its variants have been extensively used in hand-crafted feature methods for FPAD [7]. Authors of [8] used Colour texture analysis (CLBP) incorporating colour texture features, unlike other LBP variants which used only grey-scale images. Histogram of Oriented Gradient descriptors (HOG) [9], Speeded-Up Robust Features (SURF) [10], Difference of Gaussian (DoG) [11] were a few other descriptors used in FPAD. Authors of [2] used image quality features including colour diversity, specular reflection, and colour moment. Transform-based features were also used to detect PAs [12]. Manual feature engineering with high domain expertise is needed for these methods, which is one of the disadvantages of these methods.

Convolutional Neural Networks (CNN) can automatically extract distinguishable features for more effective PA detection. The recent trend in deep learning-based FPAD includes approaches such as transfer learning [13], anomaly detection, auxiliary supervision, few-shot and zero-shot learning [14] and multi-modal methods. Anomaly detection, uses only real face images for training [15], therefore any PA variant is detected as an anomaly. This approach helps to overcome dataset limitations in terms of attack variants. Auxiliary supervision [16] is also used to improve unseen attack detection. Different auxiliary features include noise, depth, and reflection. Deep CNN models mainly use RGB images, but they also use multi-modal data in NIR, SWIR, thermal and depth domains for PA detection [5]. However, multi-modal data needs extended imagery which is a less cost-effective authentication system.

Transfer learning aims to achieve unseen attack detection through domain adaptation [5] and domain generalisation [17] in FPAD. In transfer learning, either domain-specific or task-specific features were extracted for PA detection. Authors of [5] used domain-specific adaptation by fine-tuning lower layers of the pre-trained FR model using a NIR dataset in order to extract the corresponding domain-specific features for PA detection in the vehicular authentication system. A similar method was also in [18], where a multi-modal dataset was used for fine-tuning lower layers of the pre-trained FR model. The overall model included four channels. Each channel corresponded to each domain including NIR, thermal, depth and RGB. Each channel was made of a pre-trained FR model. The channels corresponding to NIR, depth and thermal, the lower layers of the pre-trained model were fine-tuned using the domain-specific data. The output features from the four were then combined to form the final feature vector and passed to the classifier layers. However, the channel corresponding to RGB data was not fine-tuned. Thus, domain-specific features were extracted from three channels in this model. Multi-modal data is used to extract domain-specific features, which requires extra imagery. It is important to note, however, that most mobile devices that use FR for authentication do not support extended imagery. It is therefore necessary to further explore RGB-based methods for obtaining discriminative features in order to enhance the FPAD. Higher layers of a deep model provide task-specific features. This concept was used in FPAD by fine-tuning top fully connected layers to extract such task-specific features [4], [13], [19]. Task-specific features were also extracted by unfreezing higher convolutional layers of deep pre-trained image classification models including VGG-19 and VGG-16 [20]. However, this method exhibited low generalisability.

Hybrid models, which combine hand-crafted features with deep features, have demonstrated improved PA detection recently [6], [21]. Authors of [22] proposed a hybrid method that utilized Discriminant Correlation Analysis (DCA), Canonical Correlation Analysis (CCA) and intensity distribution control using image contrast adjustment along with transfer learning and HOG features. Fang et al. [23] used a dual stream fusion model combining frequency, texture and semantic features. A multi-level frequency decomposition was also applied to address generalisation in this fusion method. To address the longer response times in parallel fusion methods /citedwards2021effectiveness, a serial fusion method was applied using Siamese neural networks. Sharifi [24] proposed a decision-level fusion strategy based on Log-Gabor filter features. Using the Nearest Neighbor classifier, the scores were classified. Simultaneously, feature extraction and classification were performed by a CNN model. Using the OR rule, the decisions from two modules were fused to get a final decision on the genuineness of the facial image. Cai et al [25] used a Hierarchical Fusion Module (HFM), which combined RGB image and meta-pattern instead of hand-crafted features. Hand-crafted features including colour, texture, spatial domain and frequency domain features extracted from different channel spaces were combined with deep features in [26], which proposed least square weight fusion (LSWF) of channel-

based feature classifiers. Hybrid models were therefore able to combine both hand-crafted features and deep functionality.

In recent research, vision transformers have also been used in detecting PA [27]. Vision transformers were used by fine-tuning the last fully connected layers with FAS datasets. The model exhibited improvement in cross-dataset evaluation. However, compared to the transfer learning models using deep pre-trained models, vision transformers need more computational resources. Hence, transfer learning using vision transformers is to be explored more in the FPAD context to use optimal computational parameters while achieving better generalisation.

III. Methods

Figure 2 provides a schematic diagram of the work presented in this paper. Typically, this includes utilising fine-tuning pre-trained models, hand-crafted features extraction and features fusion to form models to detect presentation attacks. To improve generalisability, deep pre-trained models were fine-tuned in order to learn task-specific features. Fusion models were also formed using features extracted from fine-tuned models and handcrafted features. Accordingly, fine-tuned and fusion models were evaluated for intra-dataset and cross-dataset performance as shown in Figure. 2. By using the SiW train set, the models were fine-tuned. The SiW, CASIA, and Replay Attack test sets were used to evaluate the performance.

A. Datasets

The models were evaluated using public FAS datasets, CASIA [28], Replay Attack [7], and SiW [29]. These datasets consist of 2D PA variants including print, photo and video attacks. Figure 1 shows samples of real and fake faces derived from three datasets. Figure 1 shows genuine facial images in the top row. In the lower row, corresponding fake facial images are displayed. A comparison of the three datasets is presented in Table.I.

Fig. 1. Real and fake facial image samples from SiW, CASIA and Replay Attack datasets. The upper row in each figure contains the real-face samples, whereas the lower row has the PA samples.

B. Fine-tuning

Existing FPAD methods used either domain-specific or task-specific features through fine-tuning deep pre-trained models in different ways. Since lower layers provide domain-specific features, some recent research followed the concept

TABLE I
Comparison of FAS datasets used in the evaluation

Dataset	**CASIA**	**Replay Attack**	**SiW**
Subject	50	50	165
Live videos	150	200	1320
Attack videos	450	1000	3300
Attack types	2 Print, Replay	Print, 2 Replay	2 print, 4 Replay
Display devices	iPad	iPhone 3GS, iPad	iPad Pro, iPhone 7, Galaxy S8, Asus MB168B

of domain-specific adaptation using multi-spectral data and a pre-trained face recognition model. On the other hand, task-specific features from RGB data were extracted by modifying and fine-tuning the classifier layers of deep pre-trained classification models. These methods showed reduced cross-dataset performance, while domain-specific adaptation required multi-spectral data. To circumvent both limitations, higher convolutional layers of the deep pre-trained classification model were fine-tuned using the SiW train set. VGG-16 and InceptionV3 were fine-tuned in a similar way.

More specifically, the fine-tuned VGG-16 and ResNet-50 models had six higher convolutional layers re-trained. The fine-tuned InceptionV3 model had eight higher convolutional layers which were retrained using the SiW dataset. The top layers included layers as follows: a fully connected layer of size 4096, batch normalization layer, dropout layer, another fully connected layer of size 4096, batch normalization layer, dropout layer, a fully connected layer of size 512, another fully connected layer of size 256 and a sigmoid layer.

C. Fusion

Fusion models using pre-trained classification models and colour texture features [6] exhibited improved intra-dataset detection performance compared to transfer learning models. Hence, fusion models were formed using fine-tuned ResNet-50 features and hand-crafted features including colour texture (CLBP), Difference of Gaussian (DoG), Histogram of Oriented Gradients (HOG) and Fast Fourier Transform (FFT). Thus, in fusion models, apart from texture features, image quality and frequency-related hand-crafted features were also combined to evaluate the performance. Fine-tuned ResNet-50 exhibited improved cross-dataset performance indicating better generalisation. Hence, ResNet-50 was selected to create fusion models with hand-crafted features. These fusion models combined task-specific features from ResNet-50 and domain-specific hand-crafted features. Fusion models were also implemented combining extracted features from fine-tuned ResNet-50, VGG-16 and InceptionV3 models. In this scenario, task-specific features from different deep models were combined and evaluated. Thus, fusion models were formed in three ways;

- Fusion of fine-tuned ResNet-50 feature and hand-crafted features.
- Fusion of fine-tuned ResNet-50 and VGG-16 features.
- Fusion of fine-tuned ResNet-50, VGG-16 and InceptionV3 features.

colour texture analysis (CLBP), Difference of Gaussian (DoG), Histogram of Oriented Gradients (HOG), and Fast Fourier Transform (FFT) features were extracted to use in

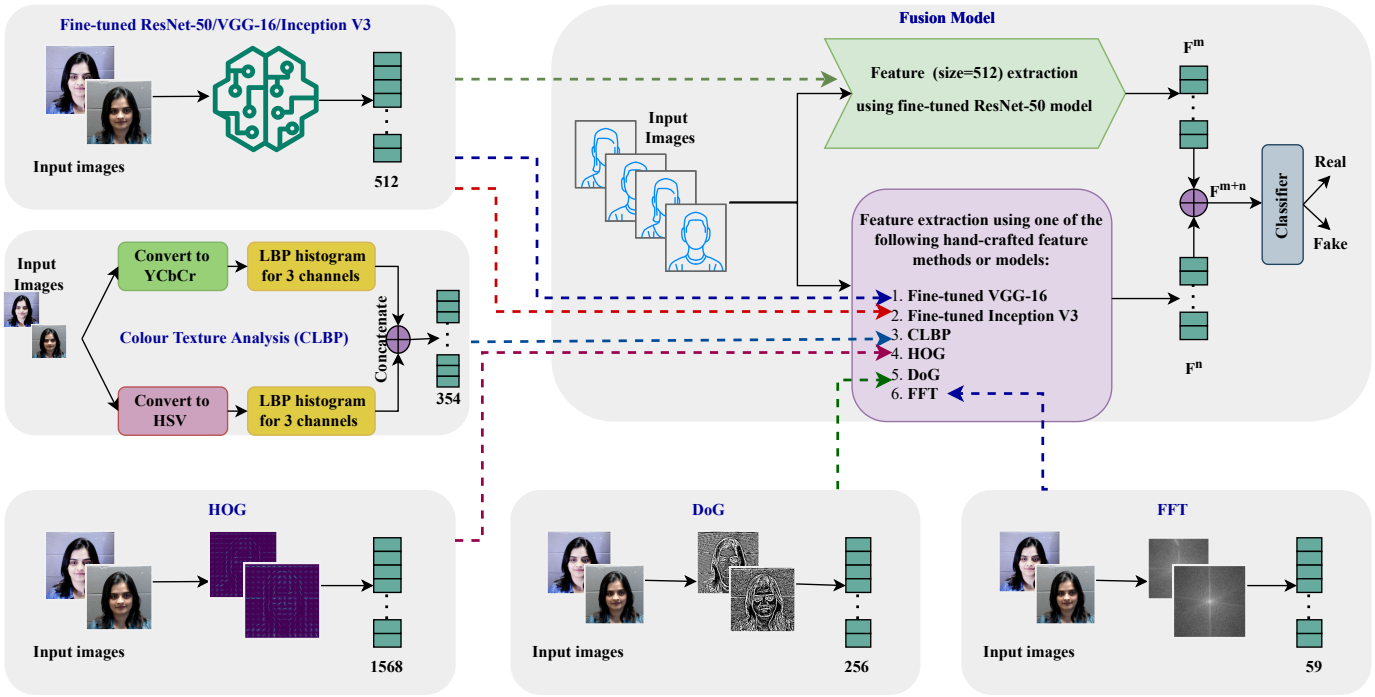


Fig. 2. A schematic of the main experiment. All deep models are pre-trained on ImageNet and fine-tuned using SiW to learn task-specific features. Both deep learning and hand-crafted features (CLBP, HOG, DoG and FFT) are fused using concatenation in combinations as detailed in the text.

fusion models. PAs introduce chrominance disparities while preserving luminance variations. Hence, the chrominance disparities cannot be identified in RGB colour space. FAS needs alternative colour spaces such as HSV and YCbCr to utilize chrominance disparities invisible in RGB colour space. HSV and YCbCr have chrominance components. As HSV and YCbCr colour spaces contain spoof-specific chrominance disparities, the images can be converted into these colour spaces and texture analysis can be performed to detect PAs [8]. The HOG provides information about the structure of the objects in the image. HOG provides edge features as well as edge direction. By extracting the edge orientation and gradients, this edge direction is provided. Thus, HOG features derived from an image represent local disparities in gradient and orientation, which can be applied to detect PAs [9]. Recapturing eradicates high-frequency features from the images, creating a disparity between real and fake facial images. These disparities can be used to identify PAs [11]. Edge detection has been used to identify differences in local features to detect PAs. DoG is applied to an image to mitigate noise and preserve high-frequency features, especially edges. Being an edge detection filter, DoG enhances the edges in the final image. The deformities in the PAs introduce differences in local features compared to the real facial image. Hence, edge detection has been used to detect recaptured images and PAs. Frequency disparities between real and fake facial images can also be extracted using FFT [12].

In colour texture analysis (CLBP) [8], RGB images were converted into HSV and YCbCr colour spaces and then the LBP of each channel in these images were extracted. LBP histograms from these six channels were combined to form a final feature vector of size 354. DoG [11] was implemented with the inner Gaussian filter standard deviation, $\sigma_i = .6$ and the outer Gaussian filter standard deviation $\sigma_j = 1$. DOG resultant images were converted to grey-scale images. Then, the histogram was extracted to get the feature vector of size 256 from the images. For HOG [9] feature extraction, the experiments used a region cell size of 16×16. Histograms from these cells together form a final feature vector of size 1568. To extract FFT [12] features, the RGB image was converted into a gray-scale and the calculated magnitude of the transform function. Since the spatial domain image of the magnitude spectrum represents the frequency features, the LBP histogram of this image was used to extract the feature vector of size 59.

The fusion models can be explained as follows: Let F^m be a feature vector with size m and F^n be another texture feature vector with size n. Thus F^{Fusion} will have the size $(m+n)$. Then the final feature vector F^{Fusion} can be represented [30] as

$$F^{Fusion} = F^m \cup F^n \tag{1}$$

Following Eq. 1, more feature vectors can be combined to form feature vectors of different sizes. Deep feature vectors from ResNet-50, VGG-16 and InceptionV3 had a size of 512. hand-crafted feature vectors had varying feature sizes. Thus, the resultant feature vector size will be the sum of the feature vectors used in the experiments. The final feature vectors were

passed to a neural network based classifier as in the fine-tuned model.

The classifier module for all models included 9 layers including four fully connected layers two batch normalization, two dropout layers and a sigmoid output layer. A detailed structure of the classifier is as follows: a fully connected layer of size 4096 followed by batch normalization and dropout, another fully connected layer of size 4096 followed by batch normalization and dropout, a fully connected layer with size 512, another fully connected layer with size 256 and a sigmoid output layer.

IV. Experiment

The experiments in this article used the SiW train set to fine-tune the deep pre-trained models. For model evaluations, SiW, CASIA and Replay Attack test sets were used. SiW test set was used for intra-dataset performance, whereas CASIA and Replay Attack test sets were for cross-dataset evaluations. Binary cross entropy loss and Adam optimizer were used for model compilation. For fine-tuning and fusion models, the learning rate used was 5×10^{-6} for all datasets. The batch size and epochs were 512 and 10 respectively.

The fine-tuned models used different scenarios. They are:

- Fine-tune the higher six convolution layers of pre-trained ResNet-50 and VGG-16 models.
- Fine-tune the higher eight convolution layers of the pre-trained InceptionV3 model.
- Fine-tune all the layers in the pre-trained ResNet-50 model, including modified top layers.
- Transfer learning using the pre-trained ResNet-50 model with modified top layers.

The fine-tuned ResNet-50 model was used for further experiments using fusion models, as it exhibited the best cross-dataset performance among the deep pre-trained models when fine-tuned. Thus, fine-tuned ResNet-50 features were combined with hand-crafted and other deep model-based features to form fusion models. The included:

- ResNet-50 features and hand-crafted features
- ResNet-50 and VGG-16 features
- ResNet-50, VGG-16 and InceptionV3 features

From the dataset videos, faces were detected from CASIA and Replay Attack frames were extracted at a rate of 2 frames per second. Using the SiW dataset, frames were extracted at 1 frame per second and face detection was performed based on the annotations provided. A random scaling of the bounding box for SiW was also performed to provide some background information and improve the diversity of facial images. The facial images from three datasets were resized to 224 $\times$ 224 pixels. The official train-test split was maintained for all three datasets. Table II summarizes the number of training and test images in each dataset.

This article reports the results using accuracy, Average Classification Error Rate (*ACER*) [1] and ROC curve analysis.

$$APCER = \frac{FP}{FP + TN}$$

TABLE II
DATASETS USED IN EXPERIMENTS AND THEIR SAMPLE SIZE IN TRAIN AND TEST PARTITIONS.

Dataset	Train			Test		
	Real	Fake	Total	Real	Fake	Total
CASIA	527	1760	2287	824	2471	3295
Replay Attack	1689	5261	6950	1928	5645	7573
SiW	14733	26057	40790	12390	22389	34779

$$BPCER = \frac{FN}{FN + TP}$$

$$ACER = \frac{APCER + BPCER}{2}$$

where *FP* is false positive, *TN* is true negative, *FN* is false negative, and *TP* is true positive.

V. Results

Intra-dataset and cross-dataset comparisons of the fine-tuned and fusion models are presented in Table. III and Table. IV respectively. The results were reported in terms of accuracy, AUC and ACER. In the table, ResNet-50 (FC)indicates, the transfer learning model using the pre-trained ResNet-50 model and ResNet-50 (ALL) is the model which had all the layers fine-tuned using SiW train set. ResNet-50, VGG-16 and InceptionV3 represent the models with fine-tuned higher convolutional as well as modified fully connected layers.

TABLE III
PA DETECTION PERFORMANCE OF MODELS IN INTRA-DATASET EVALUATION USING SIW DATASET

Models	ACC (%)	ACER(%)	AUC
ResNet-50 (FC)	97.53	2.33	0.99
ResNet-50	99.14	1.06	1.00
ResNet-50(ALL)	**99.57**	**0.51**	1.00
ResNet-50+CLBP	99.28	0.86	1.00
ResNet-50+HOG	99.27	0.90	0.99
ResNet-50+DoG	99.28	0.87	1.00
ResNet-50+FFT	99.28	0.89	0.99
ResNet-50+VGG-16	99.51	0.64	1.00
ResNet-50+Inception V3	99.23	1.01	0.99
ResNet-50 +VGG-16+ InceptionV3	99.53	0.60	1.00

From the Table. III, it is evident that the intra-dataset accuracy (99.57%) and ACER (.51%) showed as the best detection performance when all the layers of the pre-trained ResNet-50 model were fine-tuned using SiW train set. However, ResNet-50 (ALL) exhibited lower cross-dataset performance when tested with CASIA and Replay Attack, compared to fine-tuned models (ResNet-50 (FC) and ResNet-50) and fusion models (Table. IV). In the cross-dataset evaluation of CASIA, ResNet-50 showed the best performance. The model accuracy when tested with CASIA was 88.80%. The ResNet-50 model exhibited ACER of 13.98%. Nevertheless, the ResNet-50 model showed an accuracy of 85.05% and ACER of 24.61% when tested with Replay Attack.

The fusion model combining ResNet-50 and VGG-16 deep features exhibited the best cross-dataset performance (accuracy:87.43% and ACER:20.11%) with Replay Attack. Except with (ResNet-50 (FC) and ResNet-50 (ALL), cross-dataset evaluation with CASIA and Replay Attack provided accuracy

TABLE IV
CROSS-DATASET EVALUATION PERFORMANCE OF THE MODELS TRAINED ON SIW AND TESTED ON CASIA AND REPLAY ATTACK.

Models	**CASIA**			**Replay Attack**		
	ACC (%)	ACER(%)	AUC	ACC (%)	ACER(%)	AUC
ResNet-50 (FC)	75.45	48.89	0.43	74.86	48.78	0.65
ResNet-50	**88.80**	**13.98**	0.93	85.05	24.61	0.82
ResNet-50(ALL)	76.21	43.35	0.62	73.52	50.58	0.57
ResNet-50+CLBP	86.94	16.26	0.93	82.32	30.25	0.79
ResNet-50+HOG	86.65	15.47	0.91	84.19	26.62	0.77
ResNet-50+DoG	87.73	15.05	0.93	82.99	28.82	0.69
ResNet-50+FFT	87.34	16.68	0.88	82.83	29.14	0.77
ResNet-50+VGG-16	85.54	15.71	0.92	**87.43**	**20.11**	0.82
ResNet-50+Inception V3	87.39	14.20	**0.94**	85.92	23.05	0.75
ResNet-50 +VGG-16+ InceptionV3	87.00	14.21	0.92	85.60	22.80	**0.85**

greater than 80% which shows better generalisation. Fusion models slightly reduced cross-dataset performance when tested with CASIA. However, compared to ResNet-50 models, fusion models using only deep features showed an increase in performance when tested with Replay Attack. Cross-dataset performance with Replay Attack also decreased slightly with fusion models using hand-crafted features and ResNet-50 features.

ROC comparison of fine-tuned ResNet-50 models is shown in Fig. 3. The ROC analysis indicates that in intra-dataset evaluation with SiW, the models correctly detect PAs. Among the models evaluated cross-dataset, ResNet-50 with fine-tuned higher convolutional layers and fully connected layers (ResNet-50) demonstrated the highest performance. The fusion models were compared with the ResNet-50 model (Fig. 4). Compared to the ResNet-50 model, the fusion models have very similar performance, both in intra-dataset and cross-dataset evaluations. A fusion model based on the deep features of ResNet-50 and VGG-16 performed better. Fusion models, however, when formed using ResNet-50 features and hand-crafted features, showed reduced performance for cross-dataset performance in comparison to ResNet-50.

VI. DISCUSSION

PA detection evaluates the genuineness of the facial image captured by the sensor. Therefore, the PA detection task examines the disparities between fake and real facial images in terms of features such as texture, image quality and frequency in hand-crafted feature methods. In a deep learning model, lower convolutional layers learn domain-specific features and higher layers learn task-specific features. However, pre-trained image classification models cannot fully provide the features required to detect spoofing in RGB domain. The major cause is that those models were trained to detect the object in the images using the overall image features rather than checking the genuineness of the images. Hence, various pre-trained models were used after fine-tuning using FAS datasets for the PA detection task.

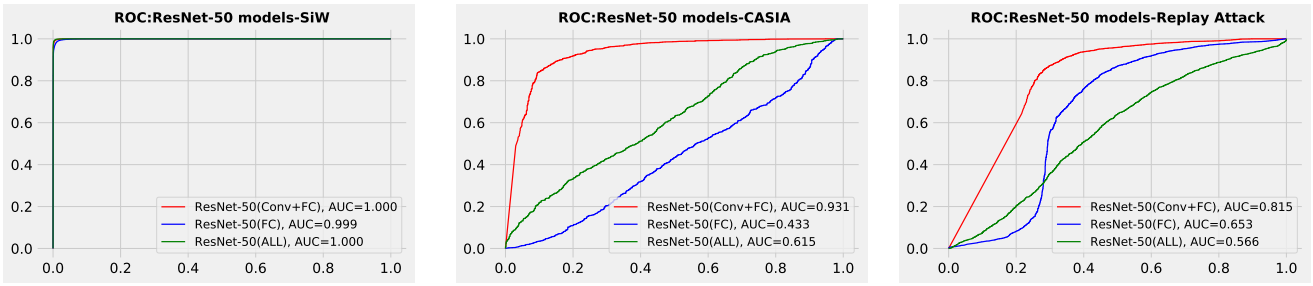
The experiments presented in this article used the pre-trained ResNet-50 model, which was fine-tuned using the FAS dataset, SiW. Fine-tuning was carried out with three different methods to analyse the performance (Section. III). Among the three methods used to fine-tune the pre-trained ResNet-50 model, the best was fine-tuning the higher convolutional layers and fully connected layers, which provided impressive performance in intra-dataset and cross-dataset evaluations as in Table. III and Table. IV.

TABLE V
COMPARISON OF THE CROSS-DATASET PERFORMANCE OF THE PROPOSED TASK-SPECIFIC LEARNING METHOD WITH SOTA METHODS. THE MODELS WERE TRAINED USING SIW DATASET AND TESTED USING CASIA.

Model	**ACER(%)**
FAS-TD-SF [31]	39.4
LGON [32]	20.56
Fusion model	14.20
ResNet-50 (Ours)	**13.98**

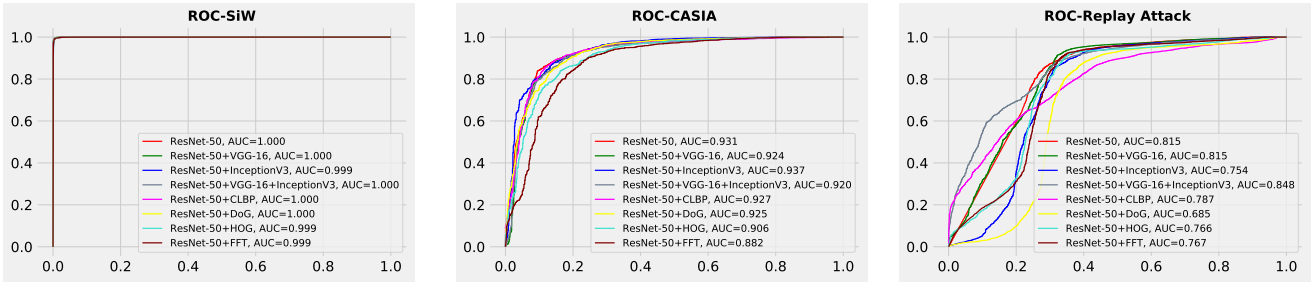
Domain-Specific Units (DSU) [5] have been used to achieve domain adaptation in PA detention fine-tuning FR models using multi-modal data. Nonetheless, higher convolutional layers provide task-specific features. Therefore the present article used a pre-trained ResNet-50 model, after fine-tuning its higher convolutional layers, and fully connected layers using only the RGB FAS dataset to extract task-specific generalisable features. This fine-tuned model exhibited performance comparable to the SOTA methods in intra-dataset as well as cross-dataset evaluations (Table. V). The fine-tuned model was also compared with fusion models, where extracted deep features from this fine-tuned model were combined with either hand-crafted features or deep features, extracted from fine-tuned VGG-16 and InceptionV3 models.

Task-specific features should be learned for better-unseen attack detection in FAS. It is a known fact that higher convolutional layers provide task-specific features. Hence, fine-tuning higher convolutional layers can enable the extraction of task-specific features which are essential for spoof detection tasks to attain generalisation. In Table. V, two SOTA methods are compared with the proposed fine-tuned and fusion models in this article. The considered methods used similar train-test dataset combinations in cross-dataset evaluation. It is evident from the cross-dataset performance ACER values that fine-tuned ResNet-50 in this article performs better compared to FAS-TD-Sf [31] and LGON [32]. However, both fine-tuned and fusion models perform slightly lower than LGON model in cross-dataset evaluation with Replay Attack.



(a) (b) (c)

Fig. 3. ROC comparison of fine-tuned ResNet-50 models trained on SiW train set and tested on (a). SiW (b). CASIA and (c). Replay Attack



(a) (b) (c)

Fig. 4. ROC comparison of fine-tuned ResNet-50 and corresponding fusion models trained on SiW train set and tested on (a). SiW (b). CASIA and (c). Replay Attack

Both CASIA and Replay Attack datasets performed better with models using deep fine-tuned features and their fusion rather than fusion models with hand-crafted features, showing that models using deep fine-tuned features are more effective in PA detection and generalisation. It also helps to avoid the disadvantages of using hand-crafted features and their extraction. Fine-tuning higher convolutional and fully connected layers of the deep pre-trained models using FAS data for FAS increases the generalisation using the inherent feature extraction capability of the deep CNN model.

VII. Conclusion

An FPAD method that fine-tunes higher convolutional layers of deep pre-trained models is presented in this article. These fine-tuned models were used for extracting deep features to form fusion models. Fusion models were formed by combining deep fine-tuned features either with hand-crafted features or with deep features from other fine-tuned models. SIW dataset was used for fine-tuning the models. CASIA and Replay Attack cross-dataset evaluations showed that fine-tuning the higher convolutional layers of the pre-trained ResNet-50 model would facilitate better task-specific feature learning. Fusion models using extracted features from fine-tuned ResNet-50, VGG-16 and Inception V3 performed better than the fusion model combining fine-tuned ResNet-50 features with hand-crafted features, in intra-dataset and cross-dataset evaluations. This illustrates that fine-tuning higher convolutional layers provide task-specific features, which in turn improves generalisation in FPAD compared to hybrid models and transfer learning. In order to test the effectiveness of this task-specific feature learning approach, more deep pre-trained models will be used after fine-tuning. In order to advance the generalisability of FPAD, future research should incorporate many datasets to facilitate more extensive intra- dataset and cross-dataset evaluations. In order to improve generalisability, the method will be evaluated after including custom losses. Feature selection and more interpretable classifiers such as gradient-boosted decision trees could also be explored in future research.

References

[1] F. Abdullakutty, E. Elyan, and P. Johnston, "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.

[2] C.-S. Fahn, C.-P. Lee, and M.-L. Wu, "A cross-dataset evaluation of anti-face-spoofing methods using random forests and convolutional neural networks," in *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, 2019, pp. 89–96.

[3] S. Kolkur and D. Kalbande, "Survey of texture based feature extraction for skin disease detection," in *2016 International Conference on ICT in Business Industry & Government (ICTBIG)*. IEEE, 2016, pp. 1–6.

[4] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *International conference image analysis and recognition*. Springer, 2017, pp. 27–34.

[5] K. Kotwal, S. Bhattacharjee, P. Abbet, Z. Mostaani, H. Wei, X. Wenkang, Z. Yaxi, and S. Marcel, "Domain-specific adaptation of cnn for detecting face presentation attacks in nir," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022.

[6] F. Abdullakutty, P. Johnston, and E. Elyan, "Fusion methods for face presentation attack detection," *Sensors*, vol. 22, no. 14, p. 5196, 2022.

[7] I. Chingovska, A. Anjos, and S. Marcel, "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, pp. 1–7.

[8] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 2636–2640.

[9] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–8.

[10] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 141–145, 2016.

[11] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3557–3560.

[12] G. Kim, S. Eum, J. K. Suhr, D. I. Kim, K. R. Park, and J. Kim, "Face liveness detection based on texture and frequency analyses," in *2012 5th IAPR international conference on biometrics (ICB)*. IEEE, 2012, pp. 67–72.

[13] C. Nagpal and S. R. Dubey, "A performance evaluation of convolutional neural networks for face anti spoofing," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[14] Y. Qin, C. Zhao, X. Zhu, Z. Wang, Z. Yu, T. Fu, F. Zhou, J. Shi, and Z. Lei, "Learning meta model for zero-and few-shot face antispoofing," *Association for Advancement of Artificial Intelligence (AAAI)*, 2020.

[15] Y. Baweja, P. Oza, P. Perera, and V. M. Patel, "Anomaly detection-based unknown face presentation attack detection," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–9.

[16] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[17] A. Costa-Pazo, D. Jiménez-Cabello, E. Vázquez-Fernández, J. L. Alba-Castro, and R. J. López-Sastre, "Generalized presentation attack detection: a face anti-spoofing evaluation proposal," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[18] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 42–55, 2019.

[19] F. Abdullakutty, E. Elyan, P. Johnston, and A. Ali-Gombe, "Deep transfer learning on the aggregated dataset for face presentation attack detection," *Cognitive computation*, pp. 1–11, 2022.

[20] S. D. Thepade, M. Dindorkar, P. Chaudhari, and S. Bang, "Face presentation attack identification optimization with adjusting convolution blocks in vgg networks," *Intelligent Systems with Applications*, vol. 16, p. 200107, 2022.

[21] L. Li, X. Feng, Z. Xia, X. Jiang, and A. Hadid, "Face spoofing detection with local binary pattern network," *Journal of visual communication and image representation*, vol. 54, pp. 182–192, 2018.

[22] M. C. Younis and H. Abuhammad, "A hybrid fusion framework to multi-modal bio metric identification," *Multimedia Tools and Applications*, pp. 1–24, 2021.

[23] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper, "Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection," *arXiv preprint arXiv:2109.07950*, 2021.

[24] O. Sharifi, "Face anti-spoofing scheme using handcraft based and deep learning methods," *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, vol. 35, no. 4, pp. 1103–1110.

[25] R. Cai, Z. Li, R. Wan, H. Li, Y. Hu, and A. C. Kot, "Learning meta pattern for face anti-spoofing," *arXiv preprint arXiv:2110.06753*, 2021.

[26] X. Song, Q. Wu, D. Yu, G. Hu, and X. Wu, "Face anti-spoofing detection using least square weight fusion of channel-based feature classifiers," EasyChair, Tech. Rep., 2020.

[27] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[28] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*. IEEE, 2012, pp. 26–31.

[29] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 389–398.

[30] W. Saad, W. A. Shalaby, M. Shokair, F. A. El-Samie, M. Dessouky, and E. Abdellatef, "Covid-19 classification using deep feature concatenation technique," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 2025–2043, 2022.

[31] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, G. Qi, J. Wan, and Z. Lei, "Exploiting temporal and depth information for multi-frame face anti-spoofing," *arXiv preprint arXiv:1811.05118*, 2018.

[32] C. Wang, B. Yu, and J. Zhou, "A learnable gradient operator for face presentation attack detection," *Pattern Recognition*, vol. 135, p. 109146, 2023.