Learning Generalizable and Identity-Discriminative Representations for Face Anti-Spoofing

XIAOGUANG TU, University of Electronic Science and Technology of China
ZHENG MA*, University of Electronic Science and Technology of China
JIAN ZHAO*, Institute of North Electronic Equipment
GUODONG DU, National University of Singapore
MEI XIE, University of Electronic Science and Technology of China
JIASHI FENG, National University of Singapore

Face anti-spoofing aims to detect presentation attack to face recognition based authentication systems. It has drawn growing attention due to the high security demand. The widely adopted CNN-based methods usually well recognize the spoofing faces when training and testing spoofing samples display similar patterns, but their performance would drop drastically on testing spoofing faces of novel patterns or unseen scenes, leading to poor generalization performance. Furthermore, almost all current methods treat face anti-spoofing as a prior step to face recognition, which prolongs the response time and makes face authentication inefficient. In this paper, we try to boost the generalizability and applicability of face anti-spoofing methods by designing a new Generalizable Face Authentication CNN (GFA-CNN) model with three novelties. First, GFA-CNN introduces a simple yet effective Total Pairwise Confusion (TPC) loss for CNN training which properly balances contributions of all the spoofing patterns for recognizing the spoofing faces. Secondly, it incorporate a Fast Domain Adaptation (FDA) component to alleviate negative effect brought by domain variation. Thirdly, it deploys the Filter Diversification Learning (FDL) to make the learned representations more adaptable to new scenes. Besides, the proposed GFA-CNN works in a multi-task manner—it performs face anti-spoofing and face recognition simultaneously. Experimental results on five popular face anti-spoofing and face recognition benchmarks show that GFA-CNN outperforms previous face anti-spoofing methods on cross-test protocol significantly and also well preserves the identity information of input face images.

CCS Concepts: •**Computing methodologies** → **Image representations; Object recognition;**

Additional Key Words and Phrases: Deep learning, computer vision, face anti-spoofing, face recognition, domain-adaptation

ACM Reference Format:
Xiaoguang Tu, Zheng Ma, Jian Zhao, Guodong Du, Mei Xie, and Jiashi Feng, 2019. Learning Generalizable and Identity-Discriminative Representations for Face Anti-Spoofing. *ACM Trans. Embedd. Comput. Syst.* 9, 6, Article 18 (November 2019), 18 pages.
DOI: 0000001.0000001

1. INTRODUCTION

As a convenient biometrics-based authentication approach, automatic face recognition has been widely adopted owing to its non-contact process and high efficiency. Despite the recent noticeable advances, the security of face recognition systems is still vulnerable to Presentation Attacks (PA) with printed photos or replayed videos. To counteract PA, lots of face anti-spoofing methods [Liu et al. 2018; Tu et al. 2019a; Tu et al. 2019a] are developed and serves as a pre-step prior to face recognition.

Most of the spoofing scenarias are produced by using multimedia devices to reproduce the identity illegally from the authorised person. During the multimedia content reproduction process, the

∗ Corresponding author: zma@uestc.edu.cn; zhaojian90@u.nus.edu
This work is partially supported by China Scholarship Council (CSC) grant 201806070011.
Author's addresses: X.G. Tu, Z. Ma, and M. Xie, School of information and communication engineering, University of Electronic Science and Technology of China, Chengdu, China; J. Zhao, Institute of North Electronic Equipment, Beijing, China; G.D.Du and J.S. Feng, Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

DOI: 0000001.0000001

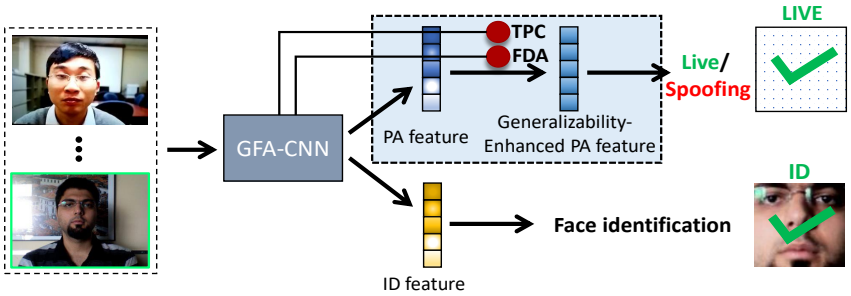

Fig. 1: Our CNN framework works in a multi-task manner, addressing face recognition and face anti-spoofing at one shot. It leverages total pairwise confusion (TPC) loss and fast domain adaption (FDA) to enhance the generalizability of the learned Presentation Attack (PA) feature and improve face anti-spoofing performance across different scenes.

cameras, display screen as well as the lighting condition are carefully tuned to obtain the reproduced content, however, which may introduce artifacts or deformations into the recaptured images/videos. To address this problem, earlier face anti-spoofing approaches adopt handcrafted features, like LBP [Chingovska et al. 2012; de Freitas Pereira et al. 2012], HoG [Komulainen et al. 2013; Yang et al. 2013] and SURF [Boulkenafet et al. 2017], to find the differences between live and spoofing faces. However, the handcrafted feature are usually designed specifically for a given dataset, which can not be well applied to new conditions [Tu et al. 2019a]. In [Yang et al. 2014], CNN was used for face anti-spoofing for the first time, with remarkable performance achieved in intra-database tests. Following their work, a number of CNN-based methods have been developed, mostly treating face anti-spoofing as a binary (live *vs*. spoofing) classification problem. However, given the limited training patterns and large variety of testing scenarios, these methods tend to suffer from overfitting and poor generalizability to new PA patterns and environments [Liu et al. 2018]. For example, the CNN model trained on dataset that contains spoofing faces recorded by iPad could not perform well on spoofing faces printed by photo, as the former PA pattern mainly involves information of light reflection, while the latter concerns more about medium materials and deformation.

In this work, we go deeper to investigate why CNN-based face anti-spoofing methods generalize poorly and attempt to improve their generalization performance and enable an anti-spoofing CNN model to be workable in various environments.

CNN-based methods differentiate live *vs.* spoofing faces by detecting certain *spoofing patterns*, including color distortion, moiré pattern, shape deformation and spoofing artifacts (*e.g.*, reflection). We find that when training CNN models, some strong patterns that contain dominant information to recognize spoofing face would take larger effect and bias the resultant model to be more discriminative for them. However, if these patterns—that the model heavily relies on—are absent in the testing data, the model performance would dramatically drop. Such overfitting to certain strong spoofing patterns is the main reason for poor generalizability of the model [Liu et al. 2018]. Apart from overfitting, domain shift [Li et al. 2018b] is also an important reason for the poor generalizability of face anti-spoofing methods. A domain here refers to a certain environment where a face image/video is captured. Different domains may differ in various factors such as illumination, background, facial appearance and camera type. Considering the huge diversity of real world environments, it is very common that different face samples are generated from different domains. For example, the domains of two paper attacks may be quite different even in case of the same face if reproduced with different pieces of paper (*e.g.* glossy *vs.* rough paper). Such domain variance may lead to distribution dissimilarity of different samples in the feature space and cause the models to fail on new domains. Furthermore, observations from prior works [RoyChowdhury et al. 2017; Graff and Ellen 2016] have shown the CNNs tend to generate near-duplicate filters over training iterations. Such near-duplicate filters may cause the CNNs mainly to focus on the dominant information that used

for classification, while ignore other information that are not important in the training dataset. The model, however, may fail on samples from a new environment where the dominant information of the training data is absence.

Based on the above observations and the analysis from previous works [Li et al. 2018b; Zhou et al. 2019; Wang et al. 2019], we propose to improve the generalizability for face anti-spoofing from three aspects: (1) alleviating overfitting on the strong spoofing patterns; (2) reducing domain shift between source and target images; (3) enhancing adaptation ability for the learned representations. Hence, we propose a new Total Pairwise Confusion (TPC) loss to balance the contributions of all involved spoofing patterns, and employ a Fast Domain Adaptation (FDA) model [Engstrom 2016] to narrow the distribution discrepancy of samples from different domains in the feature space. Furthermore, we improve the model's adaptation by enhancing the diversity of CNN convolutional filters, *i.e.,* the Filter Diversification Learning (FDL), which we show is helpful to make the model adaptive to samples from new environments. We then obtain a Generalizable Face Authentication CNN model, shorted as GFA-CNN. Since a single and unified model capable of performing face anti-spoofing and recognition is highly desired in practical systems, instead of treating face anti-spoofing as a *pre-step* prior to face recognition, for the first time, we propose to tackle the two tasks simultaneously to make face authentication more efficient, as shown in Figure 1. In the proposed framework, only the convolutional layers are shared. Face anti-spoofing and face recognition are performed separately by applying two separate fully-connected branches to transform the common face features from convolutional layers to specific features for each task.

Extensive experiments on five popular benchmarks for face anti-spoofing demonstrate the superiority of our method over the state-of-the-arts. Our code and trained models will be available upon acceptance. Our contributions are summarized as follows:

— We propose a Total Pairwise Confusion (TPC) loss to effectively relieve the overfitting problems of CNN-based face anti-spoofing models to dataset-specific spoofing patterns, which improves generalizability of face anti-spoofing methods.
— We incorporate the Fast Domain Adaptation (FDA) model to learn more robust Presentation Attack (PA) representations, which reduces domain shift in the feature space.
— We give theoretical and experimental analysis for the correlation between filter diversity and CNNs' adaptation to new environments, and propose to learn more generalizable features for face anti-spoofing by the Filter Diversification Learning.
— We develop a multi-task CNN model for face authentication, to jointly perform face anti-spoofing and face recognition, making the complete face authentication process more efficient.

2. RELATED WORK

This section reviews works closely related to our research in three fields: traditional face anti-spoofing methods, CNN-based methods and works combining anti-spoofing and authentication.

Traditional Face Anti-spoofing Methods. Most previous approaches for face anti-spoofing exploit texture differences between live and spoofing faces with pre-defined features such as LBP [Chingovska et al. 2012; de Freitas Pereira et al. 2012], HoG [Komulainen et al. 2013; Yang et al. 2013] and SURF [Boulkenafet et al. 2017], which are subsequently fed to a supervised classifier (*e.g.*, SVM, LDA) for binary classification. However, such handcrafted features are very sensitive to different illumination conditions, camera devices, specific identities, *etc*. Though noticeable performance achieved under the intra-dataset protocol, the sample from a different environment may fail the model. In order to obtain features with better generalizability, some approaches leverage temporal information, *e.g.* making use of the spontaneous motions of the live faces, such as eye-blinking [Pan et al. 2007] and lip motion [Kollreider et al. 2007]. Though these methods are effective against photo attacks, they become vulnerable when attackers simulate these motions through a paper with eye/mouth positions cut.

More recently, methods [Liu et al. 2016; Liu et al. 2018; Hernandez-Ortega et al. 2018] using the rPPG signals (*i.e.*, heart pulse signal) for face anti-spoofing are developed due to the feasibility

of detecting vital signals remotely through web-camera. In [Liu et al. 2016], Liu *et al.* develop a novel local rPPG correlation model to extract discriminative local heartbeat signal, which can be used for 3D mask face anti-spoofing regardless of the material and quality of the face mask. In [Liu et al. 2018], Liu *et al.* propose to estimate rPPG signals by sequence wise supervision, and take them as auxiliary information to guide the learning toward discriminative and generalizable cues. Obviously, the rPPG-based methods are more robust to environment changes, however, such methods need sufficient exposure rate to extract clear rPPG signals from videos, which makes it demanding for good light conditions and sensitive to different camera settings (*e.g.,* exposure rate).

Some fusion methods have also proposed to obtain a more general countermeasure against a variation of spoofing types. For example, Tronci *et al.* [Tronci et al. 2011] propose a linear fusion at frame and video level combination between static and video analysis; Schwartz *et al.* [Schwartz et al. 2011] introduce feature level fusion by using Partial Least Squares (PLS) regression based on a set of low-level feature descriptors. However, such fusion methods only focus on score or feature level, not modality level, which make the cross-domain validation not impressive due to the lack of multi-modal datasets.

CNN-based Methods. With the advent of deep learning, a natural idea is using CNNs to automatically learn highly discriminative features for face anti-spoofing. In [Yang et al. 2014], for the first time, Yang *et al.* propose to use CNNs for face anti-spoofing and achieve significant improvements on intra-test performance compared with traditional methods. After that, Xu *et al.* [Xu et al. 2015] propose to have CNNs underlying the LSTMs [Hochreiter and Schmidhuber 1997], where the local and dense property from convolutional operation could be leveraged and the temporal feature across frames can be learned and stored in LSTM units. Their method shows further improvement on the performance of intra-test protocol. Following their works, a number of CNN-based face anti-spoofing methods have been proposed [Yang et al. 2014; Li et al. 2018a; Tu et al. 2019a; Tu et al. 2019b]. However, most of them take face anti-spoofing as a binary classification problem, which makes the model suffer poor generalizability due to the overfitting to training data. Even excellent intra-test (*i.e.,*train and test within the same dataset) performance could be achieved, it seem the CNNs guess blindly when perform on cross-test protocol (*i.e.,*train and test in different datasets).

If sufficient data from various domains are available for training, the CNN-based methods could achieve satisfactory performance with fairly good generalizability. Unfortunately, current publicly available face anti-spoofing datasets are too limited to cover various potential spoofing types. To avoid the overfitting problem, Liu *et al.* [Liu et al. 2018] leverages the depth map and rPPG signal as auxiliary supervision to train CNN instead of treating face anti-spoofing as a simple binary classification problem. Their work achieves state-of-the-art cross-test performance for face anti-spoofing. In another work, the researchers regard face anti-spoofing as a de-X problem, they design a encoder-decoder architecture to inversely decompose a spoof face into a spoof noise and a live face, and then utilize the spoof noise for live/spoofing classification. Though this method could perfect avoid the overfitting issue caused by softmax loss based binary supervision, the cross-test performance is not very impressive.

Taking the cross-test as a cross domain problem, Li *et al.* [Li et al. 2018a] propose to solve such poor generalizability by using domain adaptation techniques, which aim to bridge the gap between training and testing domains. They generalize CNN to unknown conditions by minimizing the feature distribution dissimilarity across domains, *i.e.,* minimizing the Maximum Mean Discrepancy distance among representations.

Merging Anti-spoofing and Recognition. To our best knowledge, almost all previous works take face anti-spoofing as a pre-step prior to face recognition and address it as a binary classification problem. Compared with previous literature, we propose to solve face anti-spoofing and face recognition at one shot, namely the CNN model simultaneously outputs face anti-spoofing and face recognition results, which makes face authentication more intelligent and efficient. A most related work to ours is [Sajjad et al. 2018], which proposed a two-tier framework to ensure the authenticity of the user to the recognition system, namely, monitoring whether the user has passed the biometric

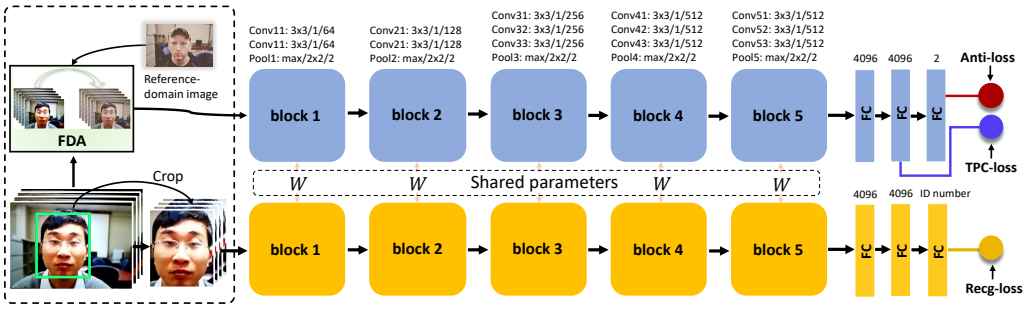

Fig. 2: Architecture of proposed GFA-CNN. The whole network contains two branches. The face anti-spoofing branch (upper) takes as input the domain-adaptive images transferred by FDA and optimized by TPC-loss and Anti-loss, while the face recognition branch (bottom) takes the cropped face images as input and is trained by minimizing Recog-loss. The structure settings are shown on top of each block. The parameter format for the convolutional layer is: filter size / stride / filter number, the number on the top of each FC layer is feature dimensions, the "ID number" indicates the number of subjects involved in training.

system as a live or spoofing one. It performs authentication based on fingerprint, palm vein print, face, *etc.*, with two separated tiers: the anti-spoofing is powered by CNN learned representations while the recognition is based on pre-defined handcrafted features like ORB points.

Different with [Sajjad et al. 2018], we build our GFA-CNN in a multi-task manner, our framework can recognize the identity of a given face, and meanwhile judge whether the face is a live or spoofing one.

3. METHOD

In this section, we first present the overall architecture of our proposed model, as illustrated in Figure 2. Then we elaborate on the details of each component.

3.1. Architecture

The proposed Generalizable Face Authentication CNN (GFA-CNN) aims to jointly address face anti-spoofing and face recognition through a multi-task architecture. This is different from existing models that mostly perform face anti-spoofing and recognition separately [Li et al. 2018a; Liu et al. 2018; Liu et al. 2016]. The proposed network has two branches: the face anti-spoofing branch and the face recognition branch. Each branch consists of 5 CNN blocks and 3 fully connected layers, and each block contains 3 CNN layers. The parameters are shared between these two branches. The common features that learned by the shared branches share similar spirit with the work DA-GAN [Zhao et al. 2017] to distinguish real and fake as well as keeping the identity information. Note that only the convolutional layers are shared in our framework, two Fully-Connected (FC) branches follow the convolutional layers, with each focuses on one specific task. Therefore, the shared conv layers contain the common feature for face images, while the two FC branches contain specific features for each task. The anti-spoofing branch takes as input raw face images with background and outputs PA features for live/spoofing classification, while the recognition branch takes cropped faces as input and gives feature for face recognition.

The "with background" images are firstly fed to convolutional layers, followed by Anti-FC layers for face anti-spoofing. Thus, the convolutional layers could learn representation containing anti-spoofing information. Then, the "cropped face" images are fed to convolutional layers, followed by Recg-FC layers for recognition. In this process, the convolutional layers mainly focus on face regions, and enable the representation to cover additional information for recognition. In the testing phase, each query image is transferred to the reference domain and then propagated forward the network.

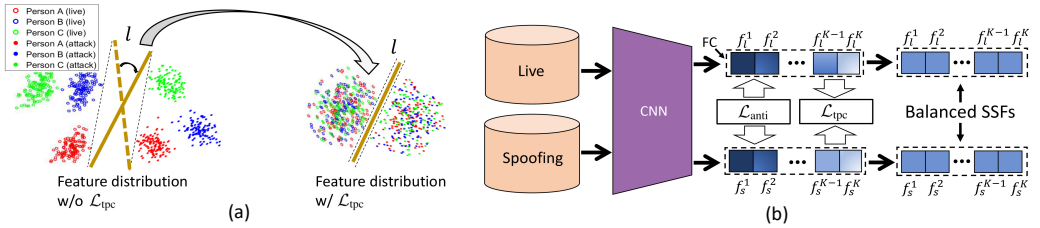

Fig. 3: (a) Visualized comparison of the learned feature distribution w/ (with) and w/o (without) $\mathcal{L}_{\text{tpc}}$. Without $\mathcal{L}_{\text{tpc}}$, the feature distribution is diverse and person-specific (left), while with $\mathcal{L}_{\text{tpc}}$, the feature distribution becomes compact and homogeneous (right). l is the classification hyperplane. (b) The contribution-balanced process of SSFs. Darker color in the FC layer indicates a higher contribution to the classification while lighter color indicates lower. Each grid represents an SSF. The trade-off game between $\mathcal{L}_{\text{tpc}}$ and $\mathcal{L}_{\text{anti}}$ can balance the contributions of SSFs to the final decision. Best viewed in color.

The overall objective function is

$$\mathcal{L} = \mathcal{L}_{\text{anti}} + \lambda_1 * \mathcal{L}_{\text{recg}} + \lambda_2 * \mathcal{L}_{\text{tpc}} + \lambda_3 * \mathcal{L}_{\text{div}}, \tag{1}$$

where $\mathcal{L}_{\text{anti}}$ is the cross entropy loss for face anti-spoofing, and $\mathcal{L}_{\text{recg}}$ is the Additive Angular Margin [Deng et al. 2019] loss for face recognition, respectively. $\mathcal{L}_{\text{tpc}}$ is the Total Pairwise Confusion (TPC) loss, $\mathcal{L}_{\text{div}}$ is the filter-diversity loss, and λ_1, λ_2 and λ_3 are the weighting parameters among different losses.

3.2. Total Pairwise Confusion Loss

In order to learn Presentation Attack (PA) representations that are adaptable to varying environment conditions, we propose a novel Total Pairwise Confusion (TPC) loss. Our inspiration comes from the pairwise confusion (PC) loss [Dubey et al. 2018] that tackles the overfitting issue in fine-grained visual classification by intentionally introducing confusion in the feature activations. For a standard classification setting with N classes, the PC loss is defined as reducing the conditional probability distance for a training pair from two different categories as $\sum_{u,v}^{n_i,n_j} ||p_\theta(\mathbf{y}|\mathbf{x}_u^i) - p_\theta(\mathbf{y}|\mathbf{x}_v^j)||$, where $p_\theta(\mathbf{y}|\cdot)$ is the conditional probability parameterized by θ, $\mathbf{x}_u^i$ and $\mathbf{x}_v^j$ are the training pair (u, v) from the class i and j respectively, and n_i and n_j are the numbers for class i and j respectively.

We modify the confusion implementation to make it applicable to the face anti-spoofing task. Our TPC loss is defined as

$$\mathcal{L}_{\text{tpc}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i \neq j}^{M} ||\psi(\mathbf{x}_i) - \psi(\mathbf{x}_j)||_2^2, \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two randomly selected images (sample pair), M is the total number of sample pairs involved in training and $\psi(\mathbf{x})$ denotes the representations of the second fully connected layer of the face anti-spoofing branch.

For a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$, our $\mathcal{L}_{\text{tpc}}$ differs from the original PC loss in two-fold: 1) TPC loss minimizes the distribution distance of a random sample pair from the training set, rather than the sample pair from two different categories, to force CNN to learn slightly less discriminative features. 2) We minimize the Euclidean distance in the feature space while the original PC loss minimizes the distance in the probability space (output of softmax) to make the samples $\mathbf{x}_i$ and $\mathbf{x}_j$ have a similar conditional probability distribution.

Our modifications are based on below considerations: 1) With face anti-spoofing taken as a binary classification issue, confusion across categories would not excessively affect the discriminability of the PA feature on differentiating live *vs*. spoofing samples. 2) Face samples related to the same subject would usually cluster in the feature space, and implementing confusion on all samples could compact and homogenize the whole feature distribution (see Figure 3 (a)), thus benefiting generalization performance. 3) As a binary classification problem of simpler structure, regularizing the

model within the feature space would be more useful than imposing regularization within the output probabilistic space.

Our $\mathcal{L}_{\text{tpc}}$ can effectively improve the generalizability of PA representations. This can be understood as follows. Suppose there are K components in the PA representations, each corresponding to one spoofing pattern, which is called a Spoof-pattern Specific Feature (SSF) in this work. As shown in Figure 3 (b), different SSFs contribute differently to the final decision. If we define the feature for a live and a spoofing sample as $F_l = (\mathbf{f}_l^1, \mathbf{f}_l^2, ..., \mathbf{f}_l^K)$ and $F_s = (\mathbf{f}_s^1, \mathbf{f}_s^2, ..., \mathbf{f}_s^K)$, respectively, where $\mathbf{f}_l^i$ is the i^{th} SSF of the live sample and $\mathbf{f}_s^i$ is the i^{th} SSF of the spoofing sample. The SSFs are ranked based on their importance to the classification of live *vs.* spoofing. On one hand, $\mathcal{L}_{\text{anti}}$ aims to enlarge the distance between F_l and F_s for better discrimination. On the other hand, $\mathcal{L}_{\text{tpc}}$ attempts to narrow the difference between F_l and F_s. As $\mathbf{f}_{l/s}^1$ contributes the most to the differentiation of live and spoofing samples, it will be impaired the most by $\mathcal{L}_{\text{tpc}}$. However, the contributions of less important SSFs, such as $\mathbf{f}_{l/s}^{K-1}$ and $\mathbf{f}_{l/s}^K$, will be enhanced by $\mathcal{L}_{\text{anti}}$ to offset the impaired discriminative ability. In this trade-off game, the contributions of all SSFs tend to be equalized, meaning more spoofing patterns are involved in the decision rather than just a couple of strong spoofing patterns specific to the training set. This could effectively alleviate overfitting risks. If some spoofing patterns disappear in testing, a fair decision can still be achieved by other patterns, ensuring CNN would not overfit to some specific features.

Actually, the feasibility of TPC loss can be explained in a more theoretical way. The fundamental reason behind overfitting is the distribution discrepancy between training and testing samples in the probability space. The "Pairwise Confusion" works to reduce the conditional probability distributions (CPD) between two inputs. For two inputs x_1 and x_2, their CPD are given by $p_\theta(y|x_1)$ and $p_\theta(y|x_2)$, respectively. The TPC works to learn θ that bring $p_\theta(y|x_1)$ and $p_\theta(y|x_2)$ closer under some distance metric, that is, make the predictions for x_1 and x_2 similar. In this way, the TPC imposes a very strong regularization over the model by adding perturbation between classification prediction over two inputs. The model is trained to minimize the classification loss even in presence of such perturbation. Thus model stability could be substantially improved. According to stability and generalization relation established in [Bousquet and Elisseeff 2002], the model could generalize better.

3.3. Fast Domain Adaptation

Besides the proposed TPC loss that balances the contribution of each spoofing pattern, we also apply fast domain adaptation (FDA) to reduce domain shift in the feature space to further improve the generalizability of our framework.

Generally, an image contains two components: content and appearance [Pan et al. 2018]. The appearance information (*e.g.*, colors, localised structures) makes up the style of images from a certain domain and is mostly represented by features in the bottom layers of CNN [Johnson et al. 2016]. For face anti-spoofing, the domain variance among face samples may introduce the distribution dissimilarity in the feature space and hurt anti-spoofing performance. Here, we employ the FDA to alleviate negative effects brought by domain changes. The FDA consists of an image transformation network $f(\cdot)$ that generates a synthetic image y from a given image x: $y = f(x)$, and a domain-loss network $\varphi(\cdot)$ that computes content reconstruction loss $\mathcal{L}_{\text{content}}$ and domain reconstruction loss $\mathcal{L}_{\text{domain}}$. The image transformation network f follows the architecture of DCGAN [Radford et al. 2015] with the pool layers replaced by strided and fractionally strided convolutions. The domain-loss network φ is a pretrained VGG-16. Training f needs a reference-domain image y_d and a dataset with lots of images (*e.g.*, COCO2014 [Lin et al. 2014] in our experiment.)

Let $\varphi_j(\cdot)$ be the j^{th} layer of $\varphi(\cdot)$ with the shape of $C_j \times H_j \times W_j$. The content reconstruction loss penalizes the output image y when it deviates in content from the input x. We thus minimize

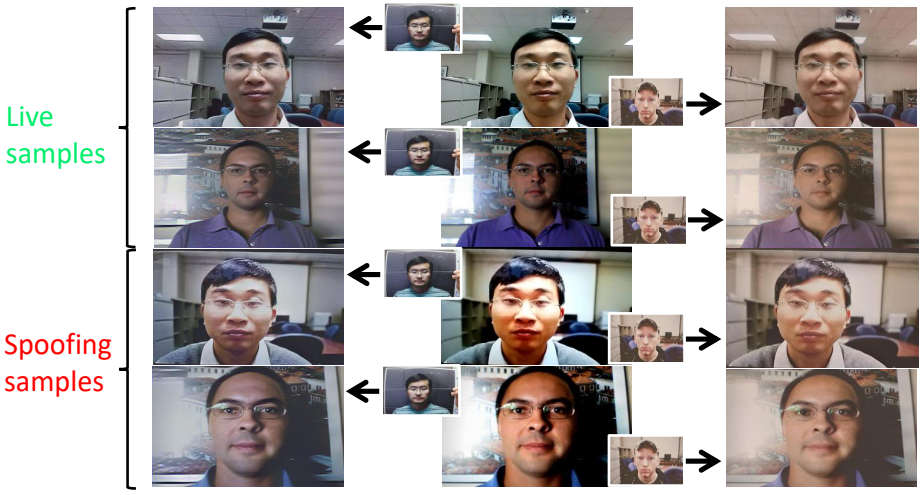

Fig. 4: Example results by FDA. The upper left and bottom right images of the images in the middle column are the reference-domain images expected to be transferred. Images of odd rows are from MSU-MFSD; images of even rows are from Replay-Attack.

the Euclidean distance between the feature representations of y and x:

$$\mathcal{L}_{\text{content}} = \frac{1}{C_j H_j W_j} ||\varphi_j(y) - \varphi_j(x)||_2^2. \tag{3}$$

The domain reconstruction loss enables the output image y to have the same domain with the reference-domain image y_d. We then minimize the squared Frobenius norm of the difference between the Gram matrices of y and y_d:

$$\mathcal{L}_{\text{domain}} = \frac{1}{C_j H_j W_j} ||G_j(y) - G_j(y_d)||_F^2. \tag{4}$$

The Gram matrix is computed by reshaping φ_j into a matrix κ, $G_j = \kappa\kappa^T / C_j H_j W_j$. Then the optimal image $\hat{y}$ is generated by solving the following objective function:

$$\hat{y} = \arg\min_P (\lambda_c \mathcal{L}_{\text{content}}(y, x) + \lambda_s \mathcal{L}_{\text{domain}}(y, y_d)), \tag{5}$$

where P is the optimal parameters of network $f(\cdot)$, x is the content image, $y = f(x)$, y_d is the reference-domain image, and λ_c and λ_s are scalars. By solving Eqn. (5), x is transferred to $\hat{y}$, preserving the content of x with the domain of y_d.

In the training of FDA, f takes as input the images from COCO2014 and is optimized by minimizing two losses, *i.e.*, the content loss $\mathcal{L}_{\text{content}}$, and the domain loss $\mathcal{L}_{\text{domain}}$ that enables y to lie in the same domain with the reference-domain image y_d. Figure 4 shows some of our domain transferred samples. The reference-domain image is sampled from the training data. Detailed analysis on the feature diversity between domains w/ and w/o FDA is provided in the ablation study section.

3.4. Filter Diversification Learning

Based on the observations in [Li et al. 2015], the CNNs always converge to a similar set of weights (filte rs). In the task of face anti-spoofing, the CNN model tends to focus all on a couple of strong spoofing patterns that can well classify spoofing and live samples during training, which makes the variance of the learned filters small. If such spoofing patterns are absent in the target environment, the CNN model would perform poorly as the learned filters don't have any domain knowledge of

other spoofing patterns. To this end, we propose the Filter Diversification Learning by enlarging the variance of convolutional filters, which we prove is effective to help CNN better adapt to new environments.

The lower layers of CNN mainly grasp general attributes such as edge, color, texture, *etc.*, however, the features of the upper layers are more dataset-specific. To this end, we only enhance the filter diversity of the upper layers of our CNN model, *i.e.,* the last CNN block. Given a set of CNN filters $W \in \mathbb{R}^{3\cdot3\times N}$, where 3 is the kernel size and N is the number of filters. We adopt the Gaussian RBF kernel to measure the similarity between the CNN filters, to ensure the learned weights are diverse from each other. The Gaussian RBF kernel has proven very effective in Euclidean spaces for a variety of kernel-based algorithms. It maps the data points to an infinite dimensional Hilbert space, which, intuitively, yields a very rich representation of the data. In $\mathbb{R}^n$, the Gaussian kernel can be expressed as:

$$K_G(x, y) = \exp(-\frac{||x - y||^2}{2\delta^2}), \tag{6}$$

which makes use of the Euclidean distance between $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$. With the Gaussian RBF kernel, we define the diversity loss as:

$$\mathcal{L}_{\text{div}} = K_G(W), \tag{7}$$

where $K_G(W)$ is the RBF function on the filter matrix W.

By minimizing $\mathcal{L}_{\text{div}}$, the Euclidean distance between each filter pair will be enlarged, so the diversity of the filters will be enhanced, making the CNN model learn features for more spoofing patterns rather than just one or two strong patterns. In this way, the CNN model trained on data from one domain will be easy to adapt to a new one.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets We evaluate GFA-CNN on five face anti-spoofing benchmarks: CASIA-FASD [Zhang et al. 2012], Replay-Attack [Chingovska et al. 2012], MSU-MFSD [Wen et al. 2015], Oulu-NPU [Boulkenafet et al. 2017] and SiW [Liu et al. 2018]. CASIA-FASD and MSU-MFSD are small datasets, containing 50 and 35 subjects, respectively. Oulu-NPU and SiW are high-resolution databases published very recently. Oulu-NPU contains 4 testing protocols: Protocol 1 evaluates the environment condition variations; Protocol 2 examines the influences of different spoofing mediums; Protocol 3 estimates the effects of different input cameras; Protocol 4 considers all the challenges above. We conduct intra-database tests on MSU-MFSD and Oulu-NPU, respectively. Cross-database tests are performed between CASIA-FASD *vs.* Replay-Attack and MSU-MFSD *vs.* Replay-Attack, respectively. The face recognition performance is evaluated on SiW, which contains 165 subjects with large variations in poses, illumination, expressions (PIE), and different distances from subject to camera. The LFW [Huang et al. 2008], the most widely used benchmark for face recognition, is also used to evaluate the face recognition performance.

Implementation Details The proposed GFA-CNN is implemented with TensorFlow [Abadi et al. 2016] framework on a single NVIDIA TITAN X GPU with 12G memory. We use Adam optimizer with a learning rate beginning at 0.0003 and decaying half after every 2,000 steps. The batch size is set as 32. λ_1, λ_2 and λ_3 in Eqn. (1) are set as 0.1, $2.5e^{-5}$ and 10, respectively. Before feeding to the face anti-spoofing branch, the training images are transferred to a reference domain by a given reference-domain image. During the training process, the two different inputs ("with background" and "cropped face" images) are fed to CNN alternatively. For training data preparation, we select one high-quality image from the training set as the reference-domain image for domain transfer, images in one mini-batch are randomly divided into pairs to calculate the TPC loss. The CNN blocks are structured the same with the convolutional part of VGG16. All experiments are performed according to the protocols provided in the datasets. Before training, the CNN blocks are first trained on the CelebFaces+ dataset [Sun et al. 2014] to obtain initial weights for face recognition. For data balance,

TPC	FDA	FDL	Intra-Test		Cross-Test	
			MFSD	Replay	MFSD → Replay	Replay → MFSD
(a) −	−	−	10.5	0.6	39.4	38.3
(b) −	+	−	11.2	0.6	36.3	34.6
(c) +	−	−	6.8	0.1	28.5	26.6
(d) +	+	−	8.3	0.3	25.8	23.5
(e) −	−	+	9.8	0.4	34.6	37.3
(f) −	+	+	10.2	0.4	36.8	33.3
(g) +	−	+	**6.4**	**0.0**	24.6	24.1
(h) +	+	+	7.4	0.3	**21.3**	**22.0**

Table I: Ablation study (HTER %). "+" means the corresponding component is used, while "-" indicates removing the component. The numbers in bold are the best results.

we triple the live samples in the training set of CASIA-FASD, MSU-MFSD and Replay-Attack with horizontal and vertical flipping.

Evaluation Metrics We have two evaluation protocols, intra-test and cross-test, which test samples from and not from the domain of the training set, respectively. To keep consistent with prior works, we report our results with the following metrics. Intra-test evaluation: Equal Error Rate (EER), Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER) and, ACER=(APCER+BPCER)/2. Cross-test evaluation: HTER, which is half of the sum of the False Rejection Rate (FRR) and the False Acceptance Rate (FAR).

4.2. Ablation Study

We first perform ablation study to reveal the role of TPC loss, FDA and FDL in our method. We retrain the proposed network by adding/ablating TPC, FDA and FDL. The results are shown in Table I. First, we consider the situation when FDL is not used: if TPC is removed, the HTERs of intra-test on MFSD increase by 2.9% (w/ FDA) and 3.7% (w/o FDA), respectively. Since Replay-Attack is usually free of severe overfitting for intra-test, it is reasonable to see the decreased HTERs are not significant when using TPC, 0.3% (w/ FDA) and 0.5% (w/o FDA) on HTER. The similar performance can also be observed when FDL is used: if TPC is used the intra-test HTERs drop by 2.8% (w/ FDA) and 3.4%(w/o FDA) on MFSD, and 0.1% (w/ FDA) and 0.4%(w/o FDA) on Replay-Attack. The experimental results indicate that TPC is effective to alleviate overfitting, no matter FAD or FDL is used or not. By comparing *Row a vs. b* and *Row c vs. d*, we observe that the intra-test performance drops slightly when FDA is used. This is reasonable, as the testing and training images are from the same domain, performing FDA to transfer testing images to the training domain could be unnecessary, which may introduce additional noise to the transferred images. We do not explore the ablation results of intra-test for FDL as these two components are not proposed to address the issues for cross-test.

For cross-test evaluation, the best result is achieved by using all the three components. If TPC is not used, the cross-test results are very poor for all the settings, see (a) (b) (e) and (f). As can be seen from (a) vs. (c) and (b) vs. (d) where FDL is not used, by using TPC the HTERs dramatically decrease by over 10% for MFSD → Replay[1], and over 11% for Replay → MFSD, no matter FDA is used or not. Similar results can also be achieved when FDL is used, with HTERs decrease at least 10% on both the two cross-test protocols regardless of the usage of FDA. The comparative results (c) vs. (d) and (g) vs. (h) demonstrate the effectiveness of FDA, where we can see FDA slightly harm the intra-test performance. This is reasonable, as the training and testing data are from the same domain for intra-test, the FDA is superfluous, which may instead bring in noise to the transformed images. However, FDA can further improve the cross-test performance, decreasing HTER by 2.7% and 3.1% on MFSD → Replay-Attack and Replay-Attack →w MFSD respectively if FDL is not

[1]The acronym $* \rightarrow \diamond$ means training on database "$*$" and testing on database "$\diamond$".

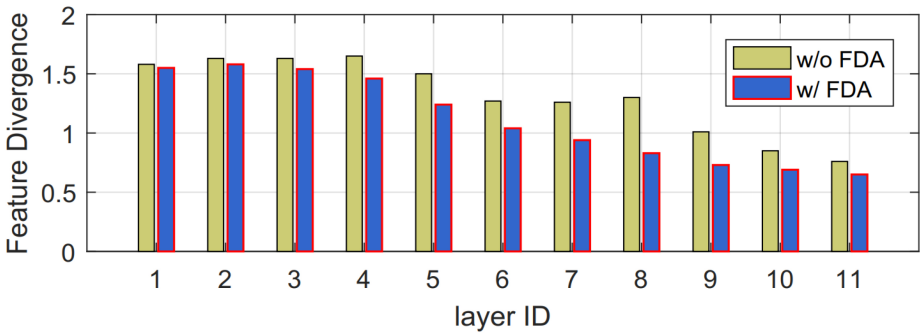

Fig. 5: Feature divergence comparison between MSU-MFSD and Replay-Attack. The numbers on x-axis correspond to the CNN layer of VGG16.

used, and 1.6% and 2.1% on MFSD → Replay-Attack and Replay-Attack → MFSD respectively if FDL is used.

To evaluate the feature diversity between domains w/ and w/o FDA, we calculate the feature divergence via symmetric KL divergence. Similar to [Pan et al. 2018], we denote the mean value of a channel from the feature embedding of CNN as F. Given a Gaussian distribution of F, with mean μ and variance σ^2, the symmetric KL divergence of this channel between domain A and B is

$$D(F_A||F_B) = KL(F_A||F_B) + KL(F_B||F_A). \tag{8}$$

$$KL(F_A||F_B) = log\frac{\sigma_A}{\sigma_B} + \frac{\sigma_A^2 + (\mu_A - \mu_B)^2}{2\mu_B^2} - \frac{1}{2}. \tag{9}$$

Denote $D(F_{iA}||F_{iB})$ as the symmetric KL divergence of the i^{th} channel. Then the average feature divergence of the layer is defined as

$$D(L_A||L_B) = \frac{1}{C}\sum_{i=1}^{C} D(F_{iA}||F_{iB}), \tag{10}$$

where C is the channel number of this layer. This metric measures the distance between the feature distributions of domain A and B. We calculate the feature divergence of each layer in a CNN model for comparison. In particular, we randomly select 5,000 face samples from MSU-MFSD and Replay-Attack, respectively. Each dataset is considered as one domain. These samples are then fed to a pre-trained VGG16 [Simonyan and Zisserman 2014] model to calculate the KL divergence at each layer following Eqn. (8). The comparison results are shown in Figure 5. As can be seen, with the FDA, the feature divergence between MSU-MFSD and Replay-Attack is significantly reduced.

In experiments (c) vs. (g) and (d) vs. (h), where only the FDL is used or removed while keeping FDA the same, TPC is essential. Referring to the results, the FDL can further improve the generalizability of our model. When only using TPC and FDA, GFA-CNN already achieves the state-of-the-art performance on MFSD → Replay-Attack (25.8%) and Reaply-Attack → MFSD (23.5%), however if FDL is used, the performance can be further improved, decreasing HTER by 2.8% on MFSD → Replay-Attack and 1.5% on Replay-Attack → MFSD, respectively. The pairs of filters are shown in Figure 6, where we can see the ones with FDL are more diverse than those without FDL, indicating the diversity on CNN weights helps to make the model better adapt to new environments.

4.3. Face Anti-spoofing Evaluation

Intra-Test We perform intra-test on MSU-MFSD and Oulu-NPU. Table II shows the comparisons of our method with other state-of-the-art methods on MSU-MFSD. For Oulu-NPU, we refer to the

Methods	EER(%)
LBP + SVM baseline	14.7
DoG + LBP + SVM baseline	23.1
IDA + SVM [Wen et al. 2015]	8.58
Color LBP [Boulkenafet et al. 2015]	10.8
GFA-CNN (ours)	6.9
Color texture [Boulkenafet et al. 2016]	4.9
Color SURF [Boulkenafet et al. 2017]	**2.2**

Table II: Intra-test results on MSU-MFSD. The numbers in bold are the best results.

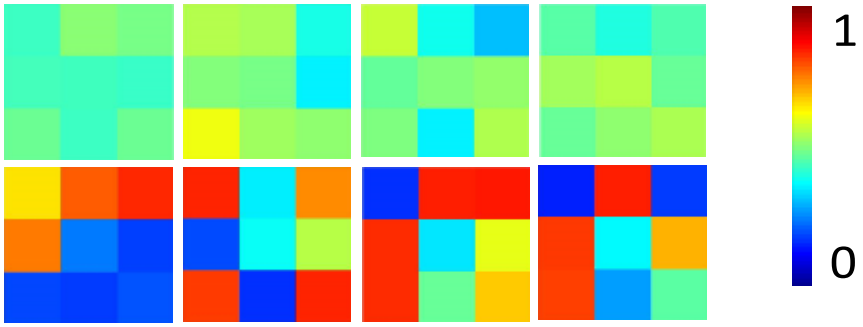

Fig. 6: Filter pairs randomly selected from our training models. The upper row illustrates filters without FDL, while the bottom are the corresponding ones with FDL.

face anti-spoofing competition results in [Boulkenafet et al. 2017] and use the best three for each protocol for comparison. All results are reported in Table III.

As shown in Table II, GFA-CNN achieves the EER of 6.9% on MFSD, ranking the 3^{rd} among all the compared methods. This result is satisfactory considering GFA-CNN is not designed blindly to pursue high performance in the intra-test setting. Actually in our experiments, we find both TPC loss and the multi-task setting may slightly decrease the intra-test performance. This is mainly because TPC loss impairs the contributions of several strongest SSFs w.r.t the training datasets. The weakening of these dataset-specific features may affect the intra-test performance, however, they can improve the performance in cross-test. For the influence of multi-task setting, we discuss this in the Discussion section. According to Table III, our method achieves the top-2 ACER in 3 out of 4 protocols on Oulu-NPU.

Cross-Test To demonstrate the strong generalizability of GFA-CNN, we perform cross-test on CASIA-FASD, Replay-Attack, and MSU-MFSD by comparing with other state-of-the-arts. We adopt the most widely used cross-test settings: CASIA-FASD *vs.* Replay-Attack and MSU-MFSD *vs.* Replay-Attack, and report comparison results in Table IV. As can be seen, GFA-CNN achieves the lowest average HTERs, and the top-2 lowest HTERs in all the cross-test settings: CASIA $\rightarrow$ Replay, Replay $\rightarrow$ CASIA, MFSD $\rightarrow$ Replay and Replay $\rightarrow$ MFSD. Especially for Replay $\rightarrow$ MFSD, GFA-CNN reduces the cross-testing HTER by 8.5% compared with the best state-of-the-art.

However, we also observe GFA-CNN has a relatively worse HTER compared with the best method on Replay Attack $\rightarrow$ CASIA-FASD. This is probably due to the "quality degradation" by FDA when the resolution of a source-domain image to be transferred is much higher than that of the reference-domain image. During the cross-testing on Replay-Attack $\rightarrow$ CASIA-FASD, the reference-domain image is selected from Replay-Attack with a low-resolution of 320×240. However, CASIA-FASD contains quite a number of images with high-resolution of 720×1280. Such

Prot.	Methods	APCER(%)	BPCER (%)	ACER (%)
1	Recod [Boulkenafet et al. 2017]	3.3	13.3	8.8
	GRADIANT [Boulkenafet et al. 2017]	1.3	12.5	6.9
	CPqD [Boulkenafet et al. 2017]	2.9	10.8	6.9
	GFA-CNN (ours)	2.5	8.9	5.7
	Auxiliary [Liu et al. 2018]	1.6	1.6	1.6
	De-spoofing[Jourabloo et al. 2018]	1.2	1.7	**1.5**
2	SZUCVI [Boulkenafet et al. 2017]	3.9	9.4	6.7
	MixedFASNet [Boulkenafet et al. 2017]	9.7	2.5	6.1
	De-spoofing[Jourabloo et al. 2018]	4.2	4.4	4.3
	Auxiliary [Liu et al. 2018]	2.7	2.7	2.7
	GRADIANT [Boulkenafet et al. 2017]	3.1	1.9	2.5
	GFA-CNN (ours)	2.5	1.3	**1.9**
3	CPqD [Boulkenafet et al. 2017]	6.8 ± 3.2	8.1 ± 5.2	7.4 ± 3.4
	MixedFASNet [Boulkenafet et al. 2017]	5.3 ± 6.7	7.8 ± 5.5	6.5 ± 4.6
	GRADIANT [Boulkenafet et al. 2017]	2.6 ± 3.9	5.0 ± 3.3	3.8 ± 2.4
	De-spoofing[Jourabloo et al. 2018]	4.0 ± 1.8	3.8 ± 1.2	3.6 ± 1.6
	GFA-CNN (ours)	3.1 ± 1.5	3.9 ± 1.3	3.5 ± 1.8
	Auxiliary [Liu et al. 2018]	2.7 ± 1.3	3.1 ± 1.7	$\mathbf{2.9} \pm 1.5$
4	Massy HNU [Boulkenafet et al. 2017]	35.8 ± 35.3	8.3 ± 4.1	22.1 ± 17.6
	CPqD [Boulkenafet et al. 2017]	32.5 ± 4.3	11.7 ± 4.8	22.1 ± 5.2
	GRADIANT [Boulkenafet et al. 2017]	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5.0
	Auxiliary [Liu et al. 2018]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	GFA-CNN (ours)	6.2 ± 4.6	9.6 ± 5.2	7.9 ± 4.8
	De-spoofing[Jourabloo et al. 2018]	5.1 ± 6.3	6.1 ± 5.1	$\mathbf{5.6} \pm 5.7$

Table III: Intra-test results on the four protocols of Oulu-NPU. The numbers in bold are the best results.

Methods	Train	Test	Train	Test	Train	Test	Train	Test	Average
	CASIA	Replay	Replay	CASIA	MFSD	Replay	Replay	MFSD	
LBP [de Freitas Pereira et al. 2013]	47.0		39.6		45.5		45.8		44.5
LBP-TOP [de Freitas Pereira et al. 2013]	49.7		60.6		46.5		47.5		51.1
Motion [de Freitas Pereira et al. 2013]	50.2		47.9		-		-		49.1
CNN [Yang et al. 2014]	48.5		45.5		37.1		48.6		44.9
Color LBP [Boulkenafet et al. 2018]	37.9		35.4		44.8		33.0		37.8
Color Tex. [Boulkenafet et al. 2018]	30.3		37.7		33.9		34.1		34.0
Color SURF [Boulkenafet et al. 2018]	26.9		**23.2**		29.7		31.8		27.9
Auxiliary [Jourabloo et al. 2018]	27.6		28.4		-		-		28.0
De-Spoof [Liu et al. 2018]	28.5		41.1		-		-		34.8
[Wang et al. 2019]	**17.5**		41.6		**5.1**		30.5		23.7
GFA-CNN (ours)	20.1		30.0		21.3		**22.0**		**23.4**

Table IV: Cross-test results (HTER %) on CASIA-FASD, Replay-Attack, and MSU-MFSD. "-" indicates the corresponding result is unavailable. The numbers in bold are the best results.

a "resolution gap" leads to a "quality degradation" of FDA which may affect the influence of face classification, as shown in the rightmost image in Figure 7.

4.4. Face Recognition Evaluation

We further evaluate the face recognition performance of our GFA-CNN on SiW and LFW. Since our method is not targeted specifically at face recognition, we only adopt the same CNN backbone (*i.e.* VGG-16) as the baseline, which we used in training our model. Face anti-spoofing is evaluated on faces of new identities that are not seen during model training while recognition is evaluated on the faces having same identities as the gallery ones, experimentally evaluating face anti-spoofing and face recognition simultaneously is impossible. For this reason, we evaluate the two tasks separately. The performance of face recognition is verified by gallery and probe strategy. On LFW, we follow the provided protocol to perform testing. On SiW we use 90 subjects for training and the other 75 subjects for testing, which is its default data splitting. This dataset also provides a frontal legacy face image corresponding to each subject. At the testing phase, we select the legacy image w.r.t

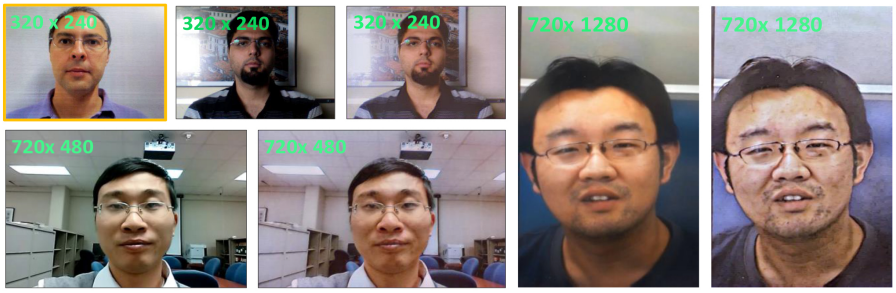

Fig. 7: Results transferred by FDA with different resolutions. The top left image is the reference-domain image. For other images of each block, the left one is the original image, and the right is the transferred image. The green number located at the top left of each image indicates the resolution.

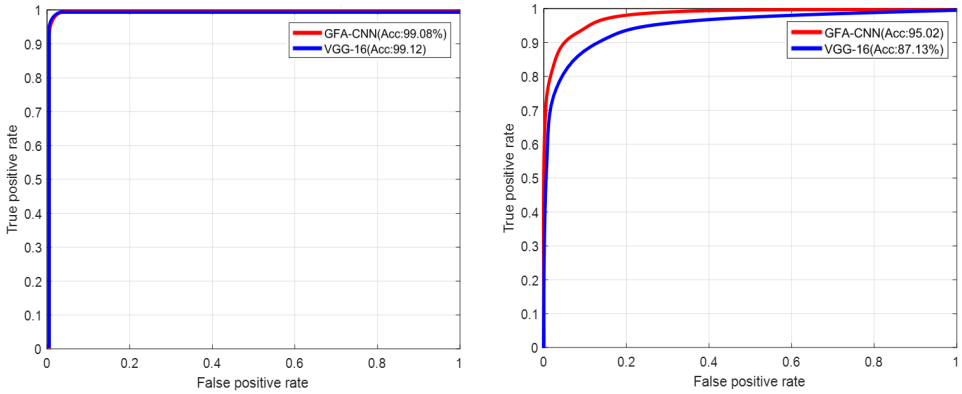

Fig. 8: Comparison of ROC curves of face verification on LFW (left) and SiW (right).

	Intra-Test		Cross-Test	
	MFSD	Replay	MFSD → Replay	Replay → MFSD
Multi-task	7.4	0.3	**21.3**	**22.0**
Single-task	**5.2**	**0.0**	26.3	25.4

Table V: The intra- and cross- test results (HTER %) by multi-task and single-task. The numbers in bold are the best results.

each subject of the testing set as the gallery faces, and use all images in the testing set (including both live and spoofing) as the probe faces.

The ROC curves of face verification are shown in Figure 8. As can be observed, GFA-CNN achieves competitive results to VGG16 on LFW, 99.08% and 99.12%, respectively. However, when testing on SiW, the declined accuracy of GFA-CNN is much lower than that of VGG16: the accuracy of GFA-CNN reduces by 4.06%, while VGG16 drops by 11.99%. The degraded performance is mainly due to face reproduction by spoofing mediums, in which some of the finer facial details might be lost. However, GFA-CNN still achieves satisfactory performance compared with VGG16. This is mainly because the face anti-spoofing and face recognition tasks mutually enhance each other, making the representations learned for face recognition less sensitive to spoofing patterns.

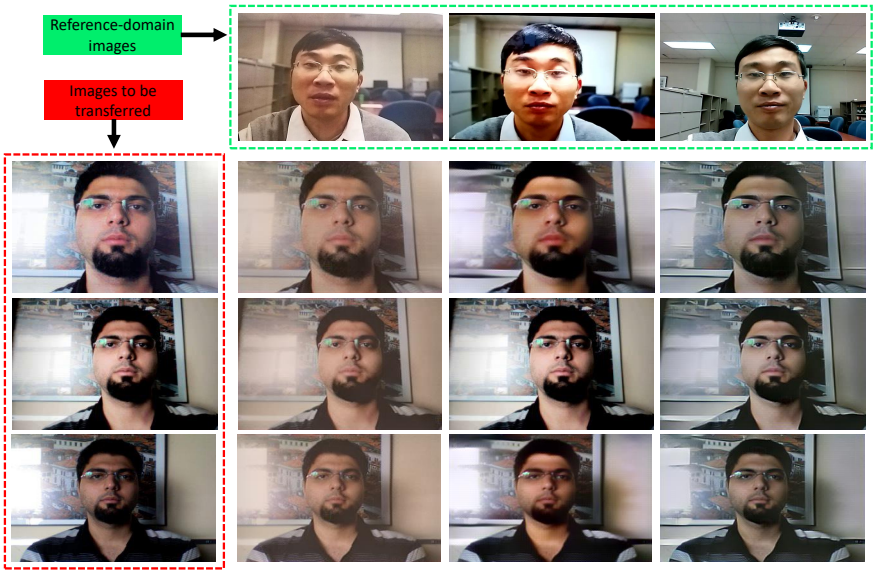

Fig. 9: The results of FDA by different reference-domain images. The images in the green bounding box are the reference-domain images from MFSD, from left to right are photo, video and live images, respectively. The images in the red bounding box are the images to be transferred (source-domain images) from Replay-Attack, from top to bottom are photo, video and live images, respectively.

	Intra-Test		Cross-Test	
	MFSD	Replay	MFSD → Replay	Replay → MFSD
–	**5.7**	**0.0**	27.8	26.4
Photo	7.4	0.3	**23.0**	**22.0**
Video	8.4	0.5	25.6	25.2
Live	8.9	0.6	25.0	24.5

Table VI: The intra- and cross- test results (HTER %) by selecting different reference-domain images. '–' indicates no reference-domain image is used. The numbers in bold are the best results.

4.5. Discussions

Multi-task Setting: We investigate how the multi-task learning affects model performance for face anti-spoofing. For this purpose, we retrain our model by single-task (*i.e.,* removing the face recognition branch), keep hyper-parameters unchanged and evaluate with the same protocol as GFA-CNN. From the results in Table V, we observe the multi-task training slightly decreases the intra-test performance of face anti-spoofing (HTER drops by 2.2% and 0.3% on MSU-MFSD and Replay-Attack, respectively). This is reasonable, since the multi-task model learns to perform two different tasks. However, the cross-test performance can be improved (HTER drops by 3.3% and 3.4% on MFSD → Replay and Replay → MFSD, respectively.) The reason behind could be that multi-task pays attention to more tasks, making the learned representation not that sample-specific, which makes the feature more generalizable.

Another two advantages for multi-task can also be observed during the training process. 1) the training becomes more stable with the Anti-loss decreasing gradually rather than dropping sharply after some steps by single-task training, suggesting multi-task setting can help alleviate overfitting. 2) as shown in Figure 8, multi-task training helps learn face representations less sensitive to spoofing patterns for face recognition. This mainly benefits from sharing parameters in the convolutional layers, giving more generic fusion features.

Reference-domain Image Selection: We discuss the effect of different selections of reference-domain image on face anti-spoofing performance. In our training, each dataset contains images form three scenarios, *i.e.,* photo, video and live. In our experiments, we randomly select one image as the reference-domain image for each scenario, the FDA results by selecting different reference-domain images are shown in Figure 9. The intra- and cross- test results by different selections are reported in Table VI, where we can see each selection considerably improves the cross-test performance. Selecting 'photo' image as the reference-domain image achieves better cross-test performance than selecting 'video' and 'live'.

However, each of the three selections has a slightly negative effect on the intra-test performance. The reason behind is, the FDA is needless for intra-test protocol as the domain between source and target domains are the same. In the transforming process, FDA may bring in additional information that harms the intra-test performance.

5. CONCLUSION

In this paper, for the first time, we present a novel CNN model to jointly address face anti-spoofing and face recognition in a mutual boosting way, which is highly desired in real applications. In order to learn more generalizable PA representations for face anti-spoofing, three novel modules are designed. 1) we propose a novel Total Pair-wise Confusion (TPC) loss to balance the contribution of each spoofing pattern, preventing the PA representations from overfitting to dataset-specific spoofing patterns; 2) we employ the Fast Domain Adaptation (FDA) technology to reduce domain shift of face samples from different environments in the feature space; 3) the Filter Diversification Learning (FDL) is used to enhance the adaptation ability of the PA representations. Extensive experiments on five popular datasets show that each of the three modules are effective to improve the generalizability of CNN based face anti-spoofing. Furthermore, the multi-task setting ensures our GFA-CNN not only achieves impressive cross-test performance for face anti-spoofing but also obtain high accuracy for face recognition, making it possible for current face authentication system to address these two tasks simultaneously.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, and others. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 265–283.

Zinelabdine Boulkenafet, Jukka Komulainen, Zahid Akhtar, Azeddine Benlamoudi, Djamel Samai, Salah Eddine Bekhouche, Abdelkrim Ouafi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Le Qin, and others. 2017. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 688–696.

Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. 2015. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2636–2640.

Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. 2016. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security* 11, 8 (2016), 1818–1830.

Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. 2017. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters* 24, 2 (2017), 141–145.

Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. 2018. On the generalization of color texture-based face anti-spoofing. *Image and Vision Computing* 77 (2018), 1–9.

Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. 2017. 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, 612–618.

Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *Journal of machine learning research* 2, Mar (2002), 499–526.

Ivana Chingovska, André Anjos, and Sébastien Marcel. 2012. 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, 1–7.

Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. 2012. LBP- TOP based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*. Springer, 121–132.

Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. 2013. Can face anti-spoofing countermeasures work in a real world scenario?. In *2013 international conference on biometrics (ICB)*. IEEE, 1–8.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4690–4699.

Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. 2018. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 70–86.

Logan Engstrom. 2016. Fast Style Transfer. https://github.com/lengstrom/fast-style-transfer/. (2016).

Casey A Graff and Jeffrey Ellen. 2016. Correlating filter diversity with convolutional neural network accuracy. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 75–80.

Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales, and Pedro Tome. 2018. Time analysis of pulse-based face anti-spoofing in visible and NIR. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 544–552.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

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

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. 2018. Face de-spoofing: Anti-spoofing via noise modeling. (2018), 290–306.

Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. 2007. Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE Transactions on Information Forensics and Security* 2, 3 (2007), 548–558.

Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. 2013. Context based face anti-spoofing. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 1–8.

Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C Kot. 2018a. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security* 13, 10 (2018), 2639–2652.

Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. 2018b. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security* 13, 7 (2018), 1794–1809.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John E Hopcroft. 2015. Convergent Learning: Do different neural networks learn the same representations?. In *FE@ NIPS*. 196–212.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 2016. 3D mask face anti-spoofing with remote photoplethysmography. In *European Conference on Computer Vision*. Springer, 85–100.

Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. 2018. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 389–398.

Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. 2007. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. (2007), 1–8.

Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. (2018), 464–479.

Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

Aruni RoyChowdhury, Prakhar Sharma, Erik Learned-Miller, and Aruni Roy. 2017. Reducing duplicate filters in deep neural networks. In *NIPS workshop on Deep Learning: Bridging Theory and Practice*, Vol. 1.

Muhammad Sajjad, Salman Khan, Tanveer Hussain, Khan Muhammad, Arun Kumar Sangaiah, Aniello Castiglione, Christian Esposito, and Sung Wook Baik. 2018. CNN-based anti-spoofing two-tier multi-factor authentication system. *Pattern Recognition Letters* (2018).

William Robson Schwartz, Anderson Rocha, and Helio Pedrini. 2011. Face spoofing detection through partial least squares and low-level descriptors. In *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 1–8.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1891–1898.

Roberto Tronci, Daniele Muntoni, Gianluca Fadda, Maurizio Pili, Nicola Sirena, Gabriele Murgia, Marco Ristori, Sardegna Ricerche, and Fabio Roli. 2011. Fusion of multiple clues for photo-attack detection in face recognition systems. In *2011 International joint conference on biometrics (IJCB)*. IEEE, 1–6.

Xiaoguang Tu, Hengsheng Zhang, Mei Xie, Yao Luo, Yuefei Zhang, and Zheng Ma. 2019a. Deep Transfer Across Domains for Face Anti-spoofing. *arXiv preprint arXiv:1901.05633* (2019).

Xiaoguang Tu, Hengsheng Zhang, Mei Xie, Yao Luo, Yuefei Zhang, and Zheng Ma. 2019b. Enhance the Motion Cues for Face Anti-Spoofing using CNN-LSTM Architecture. *arXiv preprint arXiv:1901.05635* (2019).

Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. 2019. Improving Cross-database Face Presentation Attack Detection via Adversarial Domain Adaptation. In *International Conference on Biometrics (ICB)*.

Di Wen, Hu Han, and Anil K Jain. 2015. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security* 10, 4 (2015), 746–761.

Zhenqi Xu, Shan Li, and Weihong Deng. 2015. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 141–145.

Jianwei Yang, Zhen Lei, and Stan Z Li. 2014. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601* (2014).

Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. 2013. Face liveness detection with component dependent descriptor. In *2013 International Conference on Biometrics (ICB)*. IEEE, 1–6.

Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. 2012. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*. IEEE, 26–31.

Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. 2017. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in Neural Information Processing Systems*. 66–76.

Fengshun Zhou, Chenqiang Gao, Fang Chen, Chaoyu Li, Xindou Li, Feng Yang, and Yue Zhao. 2019. Face Anti-Spoofing Based on Multi-layer Domain Adaptation. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 192–197.

Received February 2007; revised March 2009; accepted June 2009

Online Appendix to: Learning Generalizable and Identity-Discriminative Representations for Face Anti-Spoofing

XIAOGUANG TU, University of Electronic Science and Technology of China
ZHENG MA*, University of Electronic Science and Technology of China
JIAN ZHAO*, Institute of North Electronic Equipment
GUODONG DU, National University of Singapore
MEI XIE, University of Electronic Science and Technology of China
JIASHI FENG, National University of Singapore

DOI: 0000001.0000001