Multi-adversarial Discriminative Deep Domain Generalization for Face Presentation Attack Detection

Rui Shao Xiangyuan Lan Jiawei Li Pong C. Yuen
Department of Computer Science, Hong Kong Baptist University
{ruishao, jwli, pcyuen}@comp.hkbu.edu.hk, xiangyuanlan@life.hkbu.edu.hk

Abstract

Face presentation attacks have become an increasingly critical issue in the face recognition community. Many face anti-spoofing methods have been proposed, but they cannot generalize well on "unseen" attacks. This work focuses on improving the generalization ability of face anti-spoofing methods from the perspective of the domain generalization. We propose to learn a generalized feature space via a novel multi-adversarial discriminative deep domain generalization framework. In this framework, a multi-adversarial deep domain generalization is performed under a dual-force triplet-mining constraint. This ensures that the learned feature space is discriminative and shared by multiple source domains, and thus is more generalized to new face presentation attacks. An auxiliary face depth supervision is incorporated to further enhance the generalization ability. Extensive experiments on four public datasets validate the effectiveness of the proposed method.

1. Introduction

Face recognition technique has been successfully applied in a variety of applications in the real life, such as automated teller machines (ATMs), mobile phones, and entrance guard systems. The easy-access to the human face brings the convenience of face recognition, but also the presentation attacks (PA). As simple as a printed photo paper (i.e., print attack) or a digital image/video (i.e., video replay attack) could easily hack a face recognition system deployed in a mobile phone or a laptop when those spoofs are visually close to the genuine faces. Thus, how to cope with these presentation attacks prior to the step of face recognition has become an increasingly critical concern in the face recognition community.

Various face anti-spoofing methods have been proposed. Appearance-based methods aim to differentiate real and fake faces based on various appearance cues, such as color textures [5], image distortion cues [31] or deep features [32]. Temporal-based methods are proposed to extract various temporal cues, such as facial motions [23, 28, 26] or rPPG [16, 18]. Although these methods obtain promising performance in intra-dataset experiments where training and testing data are from the same dataset, the performance dramatically degrades in cross-dataset experiments where training and testing data are from different datasets. This is because existing face anti-spoofing methods capture the differentiation cues that are dataset biased [1], and thus cannot generalize well to testing data with different feature distribution compared to training data (caused by different materials of attacks or recording environments).

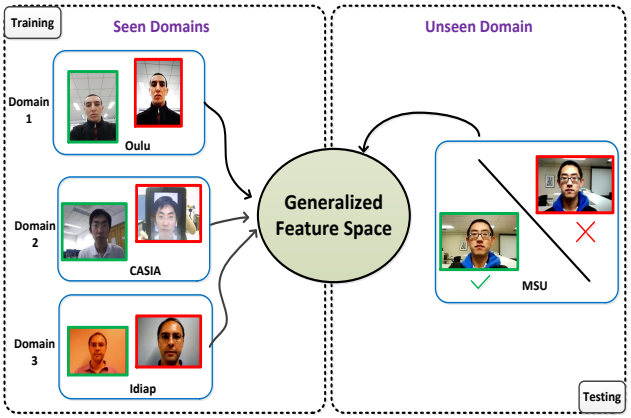

Figure 1. This paper aims to learn a feature space that is discriminative and shared by multiple source domains, and thus more generalized to new face presentation attacks.

The straightforward way to solve this problem is to exploit the domain adaptation technique [27, 12, 25, 20, 34, 7, 24, 29, 6, 30, 33, 3] to align the feature distribution between training and testing data so that the trained model with source data can be adapted on the target data. However, in the scenario of face anti-spoofing, we have no clue on test data (target domain) when we train our model. It is also difficult or impossible to collect attacks with all possible materials and in all possible environments to train and

adapt our model. To improve the generalization ability of face anti-spoofing methods without using the target domain information, this paper exploits the domain generalization approach. Domain generalization assumes that there exists a generalized feature space underlying the seen multiple source domains and the unseen but related target domain, on which a prediction model learned with training data from the seen source domains can generalize well on the unseen target domain.

The generalized feature space learned by the domain generalization approach should be shared by multiple source domains and discriminative [15, 22]. In this way, the space can exploit the common differentiation cues for face anti-spoofing across multiple source domains, which are less likely to be domain biased and thus more generalized. For example, instead of focusing on some domain-specific differentiation cues such as the screen bezel of the attack sample in CASIA dataset in Fig. 1, models learned in this generalized feature space are able to extract more generalized cues shared by all source domains. For this purpose, a multi-adversarial deep domain generalization method is proposed to automatically and adaptively learn this generalized feature space shared by multiple source domains. Specifically, under the adversarial learning scheme, the generator which is trained for producing the domain-shared features, competes with multiple domain discriminators simultaneously during the learning process, which gradually guides the learned features to be indistinguishable for multiple domain discriminators. Therefore, the feature space shared by all source domains can be automatically discovered after the feature generator fools all domain discriminators successfully. To enhance the discriminability of the learned generalized feature space during the adversarial domain generalization, we further impose a dual-force triplet-mining constraint in the learning process, which ensures the distance of each sample to its positive smaller than its negative in both intra and cross domains. Moreover, to further strengthen the generalization ability of the learned features, we incorporate face depth information as auxiliary supervision in the learning process. All of them consist of the proposed framework.

Note that a similar deep domain generalization method based on adversarial learning has been proposed in [15], which learns the generalized feature space by aligning multiple source domains to an arbitrary prior distribution via adversarial feature learning. However, simply aligning multiple source domains to a pre-defined distribution may be sub-optimal. The generalized feature space exists underlying the seen multiple source domains and the unseen target domain. This means that this generalized feature space could be learned based on the information provided by multiple source domains. To this end, we exploit the shared and discriminative information among multiple source domains to automatically and adaptively search and learn this generalized feature space without aligning any prior distribution.

2. Related Work

Face Anti-spoofing methods. Current face anti-spoofing methods can be roughly categorized into appearance-based methods and temporal-based methods. Appearance-based methods aim to detect attacks based on various appearance cues. Multi-scale LBP [19] and color textures [5] methods are proposed to extract various LBP descriptors in grayscale, RGB, HSV or YCbCr color spaces to differentiate real/fake faces. Image distortion analysis [31] detects the surface distortions caused by the lower appearance quality of images or videos compared to real face skin. Yang *et al.* [32] use CNN to extract different deep features between real and fake faces. On the other hand, temporal-based methods aim to differentiate real/fake via extracting various temporal cues through multiple frames. Dynamic textures are proposed in [23, 28, 26] to extract different facial motions. Liu *et al.* [17, 16] propose to estimate rPPG signals from RGB face videos to detect attacks. Moreover, the work proposed in [18] captures both appearance and temporal cues, which learns a CNN-RNN model to estimate the different face depth and rPPG signals between real and fake faces. However, the performance of both appearance and temporal-based methods are prone to being degraded in cross-datasets test where test data comes from different datasets (domains), and thus the feature distribution is different with that in train domain. This is due to that the above methods are likely to extract some differentiation cues that are biased to specific materials of attacks or recording environments in training datasests. Therefore, from the perspective of the domain generalization, this paper proposes to capture more generalized differentiation cues to solve both print and video replay attacks.

Deep Domain Generalization methods. Several deep domain generalization methods have been proposed. Motian *et al.* [21] propose to jointly minimize the semantic alignment loss and the separation loss on deep learning models. Li *et al.* [14] design a low-rank parameterized CNN model for end-to-end domain generalization learning. The most related work is proposed in [15], which learns a generalized feature space by aligning multiple source domains to a pre-defined distribution via adversarial learning. However, it can not be guaranteed that the pre-defined distribution is the optimal one for the feature space. Therefore, simply aligning multiple source domains to a pre-defined distribution may be sub-optimal. Instead, in our proposed deep domain generalization framework, the generalized feature space is automatically and adaptively learned based on the knowledge provided by multiple source domains.

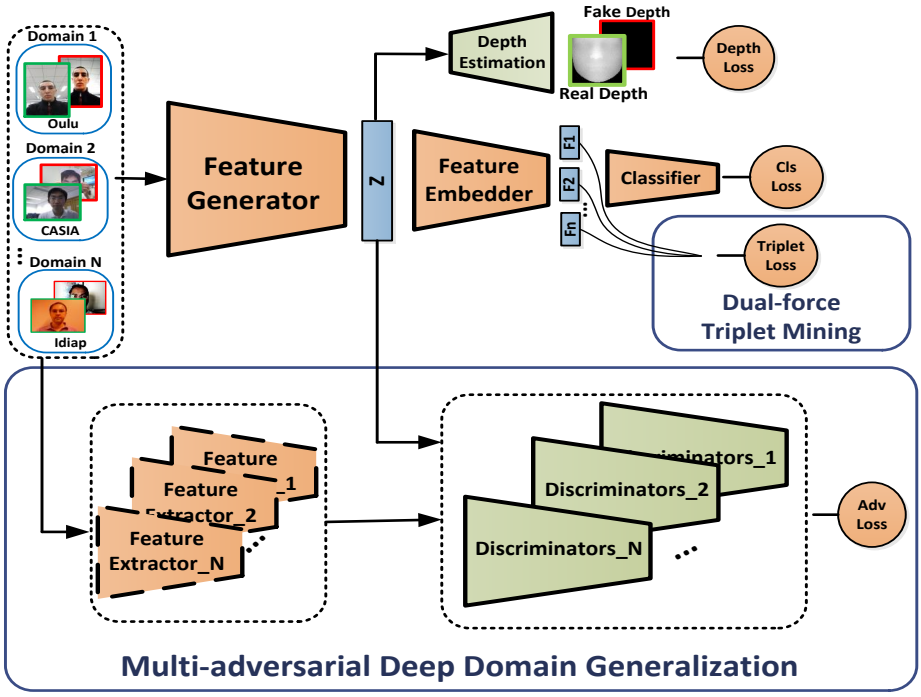

Figure 2. Overview of the proposed method. The multi-adversarial deep domain generalization is firstly proposed to learn a generalized feature space that is shared by multiple discriminative source domains. Moreover, the constraint of dual-force triplet mining is imposed on the learning process, which improves the discriminability of the learned feature space. The auxiliary face depth is further incorporated to learn more generalized differentiation cues in this feature space. The module with solid lines means it is being trained while the one with dashed lines indicates that its parameters are fixed.

3. Proposed Method

3.1. Overview

The focus of this paper is to learn a generalized feature space to cope with various unseen face presentation attacks. Although testing samples are from an unseen domain, they still share some common properties with multiple source domains in face presentation attacks. For example, the print or video replay attacks from unseen domains may be presented in different materials or under different environments compared to source domains, but they are all based on papers or video screens intrinsically. The common properties can be exploited from some shared and discriminative information across multiple source domains. That is, a feature space that is discriminative and shared by multiple source domains is more likely to be generalized well to unseen domains. Based on this idea, as illustrated in Fig. 2, this paper proposes a novel multi-adversarial discriminative deep domain generalization framework to learn this generalized feature space. Specifically, a feature generator is trained to compete against multiple domain discriminators so as to gradually learn the shared and discriminative feature space. Meanwhile, a dual-force triplet-mining constraint is imposed to improve the discrimination ability of the feature space during the adversarial learning process. Moreover, as the guidance to learn more generalized differentiation cues in the feature space, the auxiliary supervision of face depth is further incorporated in the learning process.

3.2. Multi-adversarial Deep Domain Generalization

Suppose that there are images with N source domains, denoted as $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N\}$, and corresponding labels are denoted as $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_N\}$ with K categories (K = 2 in the face anti-spoofing task where Y = 0/1 is the label of attack/real). Given the labeled data in each source domain, we may begin by exploiting the discriminative information in each source domain.

Pretrain Multiple Source Feature Extractors. For N source domains, we pre-train multiple feature extractors $(M_1, M_2, ..., M_N)$ respectively based on K-way classification with a cross-entropy loss. We take the pretraining of the feature extractor of source domain 1 as an example, which is shown as follows:

$$\mathcal{L}_{cls}(\mathbf{X}_1, \mathbf{Y}_1; M_1, C_1) = \\ -\mathbb{E}_{(x_1,y_1)\sim(\mathbf{X}_1,\mathbf{Y}_1)} \sum_{k=1}^{K} \mathbb{1}[k = y_1] log C_1(M_1(x_1)) \quad (1)$$

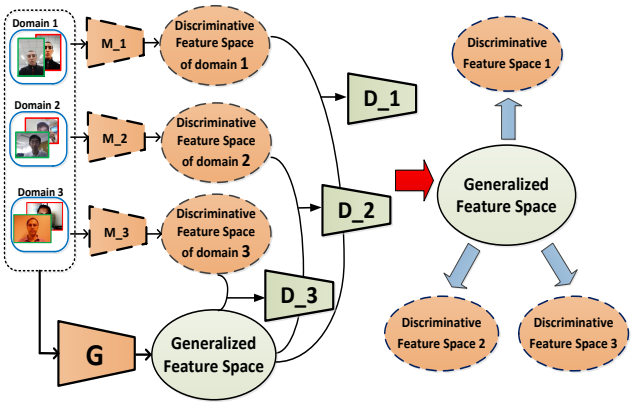

Figure 3. The details of Multi-adversarial Deep Domain Generalization. Suppose that we have three source domains for simplicity. We train one feature generator to compete with three domain discriminators simultaneously, and a shared feature space will be adaptively learned after this feature generator successfully fools all domain discriminators.

We thus obtain multiple discriminative feature spaces of source domains encoded by multiple trained feature extractors $(M_1, M_2, ..., M_N)$. However, these discriminative feature spaces contain a large portion of differentiation cues that are biased to each source domain, which disables them to be generalized well to unseen attacks.

Multi-adversarial Deep Domain Generalization. To learn a more generalized feature space for face anti-spoofing, we want to exploit the common discriminative information encoded by multiple feature extractors of source domains. More generalized differentiation cues for face anti-spoofing will thus be exploited from the common discriminative information, which are less likely to be biased to any source domain, and thus have better generalization ability.

To this end, we introduce a multi-adversarial deep domain generalization method. Because the generalized feature space contains the common discriminative information, this space can be discovered by finding the shared space of multiple discriminative source feature spaces. This means this feature space is simultaneously as similar to every discriminative feature space of source domains as possible. Suppose that we have N source domains. Accordingly, we have N discriminative feature spaces encoded by N pre-trained feature extractors respectively. N domain discriminators are introduced for N discriminative feature spaces respectively, and we train one feature generator to compete with all the N domain discriminators simultaneously. A shared feature space will thus be automatically and adaptively learned after this feature generator successfully fools all the N domain discriminators. Figure 3 shows the illustration of this multi-adversarial domain generalization process when we have three source domains. We formulate above multi-adversarial deep domain generalization as follows:

$$\begin{aligned}&\mathcal{L}_{DG}(\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N; G, D_1, D_2, ..., D_N) = \\ &\sum_{i=1}^{N} \Big(\mathbb{E}_{x\sim\mathbf{X}}[log(D_i(G(x)))] \\ &+ \mathbb{E}_{x_i\sim\mathbf{X}_i}[log(1 - D_i(M_i(x_i)))] \Big)\end{aligned} \tag{2}$$

where G denotes the feature generator, which tries to learn the generalized feature space that is indistinguishable to every discriminative source feature spaces simultaneously. D_i denotes the i-th domain discriminator that tries to distinguish the learned feature space with the discriminative feature space of source domain i. Through this multi-adversarial learning process in the feature space, the generalized feature space can be automatically learned and generated by the feature generator G.

3.3. Dual-force Triplet-mining Constraint

In print and video relay attacks, the intra-class distances are prone to being larger than the inter-class distances. Fig. 4 shows the typical condition in video replay attacks illustrating this problem. In Fig. 4, for each real subject, the fake face with the same identity has similar facial characteristics, while the real face with the different identity has different facial characteristics. This makes the negative more similar than the positive for each subject in each domain. Due to different materials of attacks or recording environments between different domains, this problem may also be severe under the cross-domain scenario. Therefore, the discrimination ability of learned generalized feature space is prone to being degraded. We thus aim to improve the discrimination ability via mining the triplet relationship among samples. Specifically, when learning the feature space, we force that: 1) the distance of each subject to its intra-domain positive smaller than to its intra-domain negative, 2) and simultaneously the distance of each subject to its cross-domain positive smaller than to its cross-domain negative. We call this as dual-force triplet-mining constraint. In this way, the discrimination ability of generalized feature space can be improved via this constraint during the domain generalization process. Therefore, we can obtain:

$$\begin{aligned}&\mathcal{L}_{Trip}(\mathbf{X}, \mathbf{Y}; G, E) = \\ &\sum_{\forall y_a=y_p, y_a\neq y_n, i=j} [\|E(G(x_i^a)) - E(G(x_j^p))\|_2^2 \\ &- \|E(G(x_i^a)) - E(G(x_j^n))\|_2^2 + \alpha_1]_+ \\ &+ \gamma \sum_{\forall y_a=y_p, y_a\neq y_n, i\neq k} [\|E(G(x_i^a)) - E(G(x_k^p))\|_2^2 \\ &- \|E(G(x_i^a)) - E(G(x_k^n))\|_2^2 + \alpha_2]_+\end{aligned} \tag{3}$$

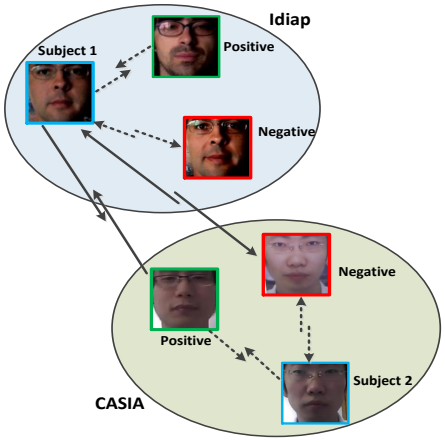

Figure 4. The illustration of Dual-force Triplet-mining Constraint. In print and video relay attacks, the negative is likely to be more similar than the positive for each subject in both intra and cross domains. This constraint attempts to solve this problem by minimizing intra-class distance while maximizing inter-class distance in both intra and cross domains.

where E denotes the feature embedder, and the superscripts a and p represent the same class, while a and n are different classes. The subscripts i and j represent the same domain, while i and k are different domains. α_1 and α_2 represent pre-defined intra-domain and cross-domain margins, respectively.

3.4. Auxiliary Face Depth Information

To exploit more generalized differentiation cues in the generalized feature space, we further incorporate the face depth cues as the auxiliary information for training our feature generator. Through the comparison on the spatial information, it can be observed that live faces have face-like depth, while faces of attacks presented in the flat and planar papers or video screens have no face depth. Therefore, the face depth information can be exploited as more generalized differentiation cues for face presentation attacks detection. We utilize the state-of-the-art dense face alignment network named PRNet [10] to estimate the depth map of real faces, which serves as the supervision for the real faces. The depth map of all zeros is set as the supervision for the fake faces. The estimated face depth information may also be domain biased. Therefore, different from the method in [18] which uses the estimated face depth to directly do classification, we incorporate the face depth as the auxiliary information into the training process of domain generalization. In this way, the feature space is guided to exploit more generalized differentiation cues related to the face depth in the learning process. This auxiliary depth information is incorporated as follows:

$$\mathcal{L}_{Dep}(\mathbf{X}; Dep) = \|Dep(G(X)) - I\|_2^2 \tag{4}$$

where Dep is the depth estimator and I is the face depth map for supervision.

3.5. Multi-adversarial Discriminative Deep Domain Generalization

As shown in Fig. 2, a classifier C is incorporated to calculated the classification loss $\mathcal{L}_{Cls}$. We formulate the objectives as mentioned above into a unified multi-adversarial discriminative deep domain generalization framework (MADDG) as follows:

$$\min_{G,E,C,Dep} \max_{D_1,D_2,...,D_N} \mathcal{L}_{MADDG} = \mathcal{L}_{DG} + \mathcal{L}_{Trip} + \mathcal{L}_{Dep} + \mathcal{L}_{Cls} \tag{5}$$

Note that due to the limited training data in face anti-spoofing datasets and the complex structure of the designed network, we decompose the training process into two phases for tractable optimization: 1) Training the G, E, C and $D_1, D_2, ..., D_N$ together, with multi-adversarial domain generalization loss, dual-force triplet-mining loss and classification loss. 2) Training the G and Dep with the auxiliary face depth information loss. The above two phases are iteratively repeated in the training process until convergence. The overall objective is to enable the feature generator G to generate the generalized feature space.

4. Experiments

4.1. Datasets

Table 1. Comparison of four experimental datasets.

Dataset	Extra light	Complex background	Attack type	Display devices
C	No	Yes	Printed photo Cut photo Replayed video	iPad
I	Yes	Yes	Printed photo Display photo Replayed video	iPhone 3GS iPad
M	No	Yes	Printed photo Replayed video	iPad Air iPhone 5S
O	Yes	No	Printed photo Display photo Replayed video	Dell 1905FP Macbook Retina

We evaluate our work on four public face anti-spoofing datasets, which contain both print and video replay attacks: Oulu-NPU [4] (O for short), CASIA-MFSD [35] (C for short), Idiap Replay-Attack [8] (I for short), and MSU-MFSD [31] (M for short). Table 1 shows the variations in these four datasets. Some samples of the genuine faces and attacks are shown in Fig. 5. From Table 1 and Fig. 5, we can see that many kinds of variations, due to the differences on materials, illumination, background, resolution and so

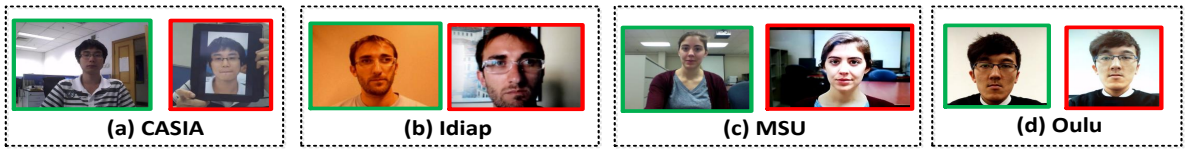

Figure 5. Sample frames from CASIA-MFSD [35], Idiap Replay-Attack [8], MSU-MFSD [31], and Oulu-NPU [4] datasets. The figures with green border represent the real faces, while the ones with red border represent the video replay attacks. From these examples, it can be seen that large cross-dataset variations due to the differences on materials, illumination, background, resolution and so on, cause significant domain shift among these datasets.

Table 2. The structure details of all components of the proposed network.

Feature Generator			**Discriminator**			**Feature Embedder & Classifier**			**Depth Estimator**		
Layer	Chan./Stri.	Out.Size	Layer	Chan./Stri.	Outp.Size	Layer	Chan./Stri.	Outp.Size	Layer	Chan./Stri.	Outp.Size
Input image			Input pool1-3			Input pool1-3			Input pool1-1+pool1-2+pool1-3		
conv1-1	64/1	256	conv2-1	128/2	16	conv3-1	128/1	32	conv4-1	128/1	32
conv1-2	128/1	256	conv2-2	256/2	8	pool2-1	-/2	16	conv4-2	64/1	32
conv1-3	196/1	256	conv2-3	512/2	4	conv3-2	256/1	16	conv4-3	1/1	32
conv1-4	128/1	256	conv2-4	1/1	3	pool2-2	-/2	8			
pool1-1	-/2	128				conv3-2	512/1	8			
conv1-5	128/1	128				Average pooling					
conv1-6	196/1	128				fc3-1	1/1	128			
conv1-7	128/1	128				fc3-2	1/1	1			
pool1-2	-/2	64									
conv1-8	128/1	64									
conv1-9	196/1	64									
conv1-10	128/1	64									
pool1-3	-/2	32									

on, exist across these four datasets. Therefore, significant domain shift exists among these datasets.

4.2. Experimental Setting

We regard one dataset as one domain in our experiment. For simplicity, three datasets in four are randomly selected as source domains where we conduct the domain generalization, and the remaining one is the unseen domain for testing, which cannot be accessed in the training process. Half Total Error Rate (HTER) [2] (half of the summation of false acceptance rate and false rejection rate) and Area Under Curve (AUC) are used as the evaluation metrics in our experiments.

4.3. Implementation Details

Network Structure. Our deep network is implemented on the platform of PyTorch. The detailed structure of the proposed network is illustrated in Table 2. To be specific, each convolutional layer in the feature generator, feature embedder and depth estimator is followed by a batch normalization layer and a rectified linear unit (ReLU) activation function, and all convolutional kernel size is 3×3. Following the standard setting in [14], each convolutional layer in the discriminator is followed by a batch normalization layer and a LeakyReLU activation function, and all kernel size is 4×4. The size of input image is $256 \times 256 \times 6$, where we extract the RGB and HSV channels of each input image. Inspired by the residual network [11], we use a short-cut connection, which is concatenating the responses of pool1-1, pool1-2 and pool1-3, and sending them to conv4-1 for depth estimation. This operation helps to ease the training procedure, and enables the auxiliary information of face depth simultaneously to affect different layers of the feature generator in the learning process.

Training Details. The Adam optimizer [13] is used for the optimization. As described in section 3.5, we train the whole network with two iterative phases. Due to different model complexity between the two training phases, we use the learning rate 1e-5 in the first phase, which trains the G, E, C and $D_1, D_2, ..., D_N$ together. G and Dep are trained in the second phase with the learning rate 1e-4. The batch size is 20 per domain, and thus 60 for 3 training domains totally. The hyperparameters γ, α_1, and α_2 are set to 0.1, 0.1, and 0.5, respectively.

Testing. For a new testing sample x, its classification score l is calculated for testing as follows: $l = C(E(G(x)))$ where G, E, C are the trained feature generator, feature embedder, and classifier, respectively.

4.4. Experimental Comparison

4.4.1 Baseline Methods

We compare several state-of-the-art face anti-spoofing methods as follows: **Multi-Scale LBP (MS_LBP)** [19] ; **Binary CNN** [32]; **Image Distortion Analysis (IDA)** [31]; **Color Texture (CT)** [5]; **LBPTOP** [23]; and **Auxiliary** [18]: This method learns a CNN-RNN model to estimate the face depth from one frame and rPPG signals through multiple frames. To fairly compare our method only using one frame information, we implement its face depth

Table 3. Comparison to face anti-spoofing methods on four testing sets for domain generalization on face anti-spoofing.

Method	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MS_LBP	29.76	78.50	54.28	44.98	50.30	51.64	50.29	49.31
Binary CNN	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
IDA	66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
Color Texture	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
LBPTOP	36.90	70.80	42.60	61.05	49.45	49.54	53.15	44.09
Auxiliary(Depth Only)	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
Auxiliary(All)	–	–	28.4	–	27.6	–	–	–
Ours (MADDG)	**17.69**	**88.06**	**24.5**	**84.51**	**22.19**	**84.99**	**27.98**	**80.02**

Table 4. Comparison to adversarial domain generalization method on four testing sets for domain generalization on face anti-spoofing.

Method	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MMD-AAE	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
Ours (MADDG)	**17.69**	**88.06**	**24.5**	**84.51**	**22.19**	**84.99**	**27.98**	**80.02**

estimation component(denoted as Auxiliary(Depth Only)). We also compare its reported results (denoted as Auxiliary(All)). Moreover, we also compare the related state-of-the-art method in domain generalization for the face anti-spoofing task: **MMD-AAE** [15].

4.4.2 Comparison Results

From the comparison results in Table 3 and Fig. 6, it can be seen that the proposed method performs better than all the state-of-the-art face anti-spoofing methods [19, 32, 31, 5, 18]. This is due to that all existing face anti-spoofing methods focus on learning a feature space from multiple source domains that only fits to data in the source domains. Comparatively, the proposed multi-adversarial discriminative deep domain generalization explicitly exploits the domain relationship of multiple source feature spaces, and learns the shared and discriminative information between them. This learns a generalized feature space that is more likely to be shared between source domains and unseen target domain, and thus it is more able to extract more generalized differentiation cues for face anti-spoofing.

Moreover, in Table 4 and Fig. 6, compared to the state-of-the-art domain generalization method [15], we also outperform it for the face anti-spoofing task. This illustrates that comparing to the feature space learned by aligning multiple source domains to a pre-defined distribution, the feature space that is automatically and adaptively learned by our proposed domain generalization framework is more feasible for the task of face anti-spoofing.

4.5. Discussion

4.5.1 Ablation Study

The experimental results of ablation study for all testing sets are shown in Table 5. **MADDG** denotes the proposed framework. **MADDG_wo/mgan** denotes that the proposed network without the multi-adversarial domain generalization component. In this setting, we remove the multiple domain discriminators $(D_1, ..., D_N)$ in our network in the training process. **MADDG_wo/trip** denotes that the proposed network without the dual-force triplet-mining constraint component. In this setting, we do not calculate and backpropagate the dual-force triplet-mining loss in the training process. **MADDG_wo/dep** denotes that the proposed network without incorporating auxiliary face depth information. In this setting, we remove the depth estimator Dep in the training process.

Table 5 shows that the performance of the proposed network degrade if any component is excluded. This verifies the contribution of each component to the whole network, and shows that the proposed network optimizing all components simultaneously in a unified framework can obtain much better performance.

4.5.2 Fusion strategies comparison

Fusion strategies are usually utilized when we have multiple domains data. Thus, we added two more baselines for comparison, namely score-level fusion and feature-level fusion in Table 6. In score-level fusion, we train multiple AlexNets for all source domains respectively, and use average fusion on the testing scores of all trained CNNs on the target domain. In feature-level fusion, like [9], we train multiple AlexNets and fuse the features from FC7 layers by concatenation. One more fully connected layer is integrated to classify the fused feature. Table 6 shows our method outperforms the above two kinds of fusion strategies. Simple fusion strategies cannot cope with various cross-domain scenarios so that in some scenarios such as O&M&I to C, the performance of both baseline methods drop significantly. Comparatively, our method is robust in all scenarios.

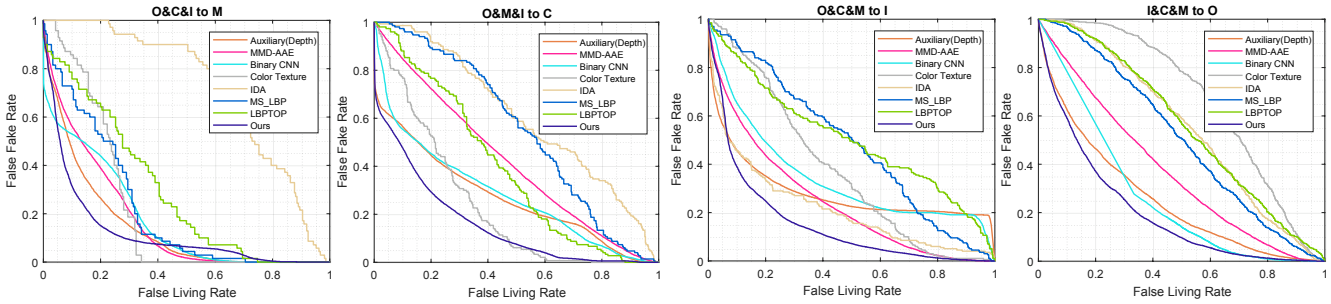

Figure 6. ROC curves of four testing sets for domain generalization on face anti-spoofing.

Table 5. Evaluation of different components of the proposed framework on four testing sets for domain generalization on face anti-spoofing.

Method	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MADDG_wo/mgan	21.55	85.83	28.67	82.27	36.50	63.15	29.63	77.40
MADDG_wo/trip	20.84	85.95	30.46	77.99	34.99	71.37	29.75	75.93
MADDG_wo/dep	34.29	69.92	39.95	62.42	37.44	62.82	39.39	64.19
Ours(MADDG)	**17.69**	**88.06**	**24.5**	**84.51**	**22.19**	**84.99**	**27.98**	**80.02**

Table 6. Comparison to fusion strategies on four testing sets for domain generalization on face anti-spoofing.

Method	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
Score_Fusion	21.00	86.18	46.62	57.05	34.17	71.53	31.12	76.42
Feature_Fusion	25.62	74.57	52.32	48.23	46.29	52.71	32.56	76.01
Ours (MADDG)	**17.69**	**88.06**	**24.5**	**84.51**	**22.19**	**84.99**	**27.98**	**80.02**

4.5.3 Limited source domains

We evaluate the domain generalization ability of the proposed method when extremely limited source domain datasets are available (i.e. only two source datasets). Since significant domain variation exists between **MSU** and **Idiap** datasets, we choose these two datasets as source domains. As such, the remaining ones (**Oulu** and **CASIA**) are chosen for testing. The results in Table 7 show the proposed method performs better than other methods. This verifies our method is more effective even in the challenging case.

Table 7. Comparison of domain generalization with limited source domains for face anti-spoofing.

Method	M&I to C		M&I to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MS_LBP	51.16	52.09	43.63	58.07
IDA	45.16	58.8	54.52	42.17
CT	55.17	46.89	53.31	45.16
LBPTOP	45.27	54.88	47.26	50.21
Ours	**41.02**	**64.33**	**39.35**	**65.10**

Moreover, compared to the results in Table 7, when we have more source domains such as the normal setting of domain generalization in Table 3, the other methods cannot get much improvement and the proposed method outperforms them in a larger gap. This means our method is more able to exploit domain shared and discriminative properties to learn more generalized cues when more source domains are available, and thus the advantage of domain generalization can be better exploited by our method.

5. Conclusion

To improve the generalization ability for face anti-spoofing, this paper exploits the technique of domain generalization to learn a generalized feature space without using target domain data. Specifically, a novel multi-adversarial deep domain generalization method is proposed to train one feature generator to compete with multiple domain discriminators simultaneously, so that the generalized feature space can be automatically and adaptively learned. The discriminability of the generalized feature space is improved by a dual-force triplet-mining constraint in the feature learning process. Meanwhile, the face depth supervision is incorporated to further enhance the generalization ability of this feature space. Extensive experiments among four public datasets validate the effectiveness of the proposed method.

6. Acknowledgement

This project is partially supported by Hong Kong RGC GRF HKBU12201215. The work of X. Lan is partially supported by HKBU Tier 1 Start-up Grant.

References

[1] Torralba. Antonio and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1

[2] Samy Bengio and Johnny Mariéthoz. A statistical significance test for person authentication. In *The Speaker and Language Recognition Workshop*, 2004. 6

[3] Bharath Bhushan Damodaran, Benjamin Kellenberger, Remi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, 2018. 1

[4] Zinelabinde Boulkenafet and et al. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FG*, 2017. 5, 6

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. In *IEEE Trans. Inf. Forens. Security, 11(8): 1818-1830*, 2016. 1, 2, 6, 7

[6] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017. 1

[7] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *CVPR*, 2018. 1

[8] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012. 5, 6

[9] Litong Feng and et al. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *Journal of Visual Communication and Image Representation*, 38:451–460, 2016. 7

[10] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[12] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Duplex generative adversarial network for unsupervised domain adaptation. In *CVPR*, 2018. 1

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 2, 6

[15] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 2, 7

[16] Siqi Liu, Xiangyuan Lan, and Pong C. Yuen. Remote photoplethysmography correspondence feature for 3D mask face presentation attack detection. In *ECCV*, 2018. 1, 2

[17] Siqi Liu, Pong C. Yuen, Shengping Zhang, and Guoying Zhao. 3D mask face anti-spoofing with remote photoplethysmography. In *ECCV*, 2016. 2

[18] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018. 1, 2, 5, 6, 7

[19] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IJCB*, 2011. 2, 6, 7

[20] Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bul, Barbara Caputo, and Elisa Ricci. Boosting domain adaptation by discovering latent domains. In *CVPR*, 2018. 1

[21] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 2

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. In *TKDE*, 2010. 2

[23] Tiago Freitas Pereira and et al. Face liveness detection using dynamic texture. In *EURASIP Journal on Image and Video Processing, (1): 1-15*, 2014. 1, 2, 6

[24] Pedro O. Pinheiro. Unsupervised domain adaptation with similarity learning. In *CVPR*, 2018. 1

[25] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 1

[26] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3D mask face anti-spoofing. In *IJCB*, 2017. 1, 2

[27] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Feature constrained by pixel: Hierarchical adversarial deep domain adaptation. In *ACM MM*, 2018. 1

[28] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Joint discriminative learning of deep dynamic textures for 3D mask face anti-spoofing. In *IEEE Trans. Inf. Forens. Security, 14(4): 923-938*, 2019. 1, 2

[29] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 1

[30] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *CVPR*, 2018. 1

[31] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. In *IEEE Trans. Inf. Forens. Security, 10(4): 746-761*, 2015. 1, 2, 5, 6, 7

[32] Jianwei Yang, Zhen Lei, and Stan Z. Li. Learn convolutional neural network for face anti-spoofing. In *arXiv preprint arXiv:1408.5601*, 2014. 1, 2, 6, 7

[33] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *CVPR*, 2018. 1

[34] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, 2018. 1

[35] Zhiwei Zhang and et al. A face antispoofing database with diverse attacks. In *ICB*, 2012. 5, 6