This CVPR 2020 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Domain Agnostic Feature Learning for Image and Video Based Face Anti-spoofing

Suman Saha
ETH Zurich
suman.saha@vision.ee.ethz.ch

Wenhao Xu
ETH Zurich
wenhxu@student.ethz.ch

Menelaos Kanakis
ETH Zurich
menelaos.kanakis@vision.ee.ethz.ch

Stamatios Georgoulis
ETH Zurich
stamatios.georgoulis@vision.ee.ethz.ch

Yuhua Chen
ETH Zurich
yuhua.chen@vision.ee.ethz.ch

Danda Pani Paudel
ETH Zurich
paudel@vision.ee.ethz.ch

Luc Van Gool
KU Leuven & ETH Zurich
vangool@vision.ee.ethz.ch

Abstract

Nowadays, the increasingly growing number of mobile and computing devices has led to a demand for safer user authentication systems. Face anti-spoofing is a measure towards this direction for biometric user authentication, and in particular face recognition, that tries to prevent spoof attacks. The state-of-the-art anti-spoofing techniques leverage the ability of deep neural networks to learn discriminative features, based on cues from the training set images or video samples, in an effort to detect spoof attacks. However, due to the particular nature of the problem, i.e. large variability due to factors like different backgrounds, lighting conditions, camera resolutions, spoof materials, etc., these techniques typically fail to generalize to new samples. In this paper, we explicitly tackle this problem and propose a class-conditional domain discriminator module, that, coupled with a gradient reversal layer, tries to generate live and spoof features that are discriminative, but at the same time robust against the aforementioned variability factors. Extensive experimental analysis shows the effectiveness of the proposed method over existing image- and video-based anti-spoofing techniques, both in terms of numerical improvement as well as when visualizing the learned features.

1. Introduction

Increasingly, people use computing devices, such as laptops and smartphones, to work, pay their bills, purchase things as well as interact with their social circle, entertain themselves, *etc*. Given the constant use we make of these devices, it is important to develop convenient, yet secure, ways to log into them. Lately, biometric authentication, and in particular face recognition, has emerged as an attractive way of user identification due to the unique nature of each individual's face in combination with the ease-of-use of this approach (e.g. Apple's FaceID). At the same time, however, hackers have become more inventive in their attempts to spoof someone's face in order to fool the authentication system. Typical examples include printing one's face picture on paper (print attack), playing a video depicting the person's face on another device (replay attack), wearing a special mask to closely imitate someone's facial appearance (mask attack), *etc*. Understandably, being able to effectively detect such attacks, formally known as face anti-spoofing

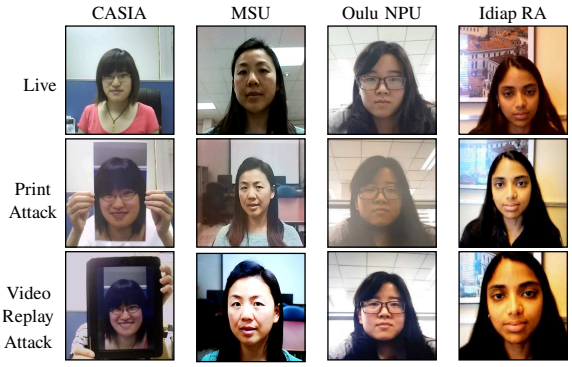


Figure 1. Sample frames from the four publicly available face anti-spoofing datasets: CASIA-MFSD [52], MSU-MFSD [48], Oulu-NPU [8] and Idiap Replay-Attack (RA) [10]. Note that, a large variability can be observed due to factors like different attack instruments, backgrounds, lighting conditions, camera resolutions etc. resulting significant domain shift among these datasets.

(FAS), is a critical problem in computer vision.

On the one hand, traditional approaches to face anti-spoofing rely on hand-crafted features, like LBP [10], HoG [25] and SURF [7], to detect differences in texture between the live and spoof face images, or heuristics, like eyeblink [37] and lip motion [24], to identify regularities that are absent from the spoof attacks. However, the aforementioned methods are either not applicable to all possible spoof attacks, *i.e*. print, replay, and mask, or they fail to generalize to different datasets, since the learned features specialize to the 'trained' textures, which largely vary between datasets due to factors like different backgrounds, lighting conditions, camera resolutions, spoof materials, *etc*. as can be seen in Fig. 1.

On the other hand, modern approaches use convolutional neural networks (CNNs) [33, 22] that have shown impressive performance in many computer vision tasks, largely attributed to the great representational power of their learned features when trained on large-scale datasets. Despite the improved performance, there are still open challenges in FAS. A notable one is the domain[1] shift [30] problem. The latter occurs when a network trained on one dataset (source domain) is tested on a completely unseen dataset (target domain). This is referred to as "cross-testing" in the FAS literature, while training and testing on the same dataset is referred to as "intra-testing". The existing deep learning based approaches show promising results for intra-testing, but their performance dramatically degrades when evaluated under a cross-testing setup [40]. The main reason for this performance drop is the feature distribution dissimilarity (see Fig. 2) between the source and target domains caused by several dataset specific cues, such as differences in: (1) environmental conditions (illumination, background), (2) spoofing mediums (printers, display screens), and (3) the quality of video capturing devices (different mobile phones, tablets). Thus, a model learns to differentiate between live and spoof samples based on these dataset dependent cues, but fails to correctly classify samples from unknown datasets having different sets of cues.

In this paper, we address the aforementioned domain shift problem in FAS under the domain generalization setting. That is, the network is trained on multiple datasets (source domains), but then tested on a completely unseen dataset (target domain). Our goal is to generate domain agnostic feature representations using the source domain samples that would generalize to the unseen target domain samples, so that each sample, regardless of its domain origin, can effectively be classified as live or spoof. To this end, we propose the use of class-conditional domain discriminator modules coupled with a gradient reversal layer [15]. The former take the feature representations generated from a backbone network, and try to classify from which source domain each sample comes, conditioned on the class it belongs (i.e. live or spoof). The latter acts as an identity transform during the forward propagation, but multiplies the gradient by a certain negative constant during the backward propagation, essentially reversing the objective of its subsequent layers. In our case, this practically means that the backbone network is now tasked with the extra objective of generating live and spoof feature representations that are indistinguishable across domains. Note that, our method works for both image-based and video-based inputs, but we explicitly avoid to include extra components as input, like depth or rPPG signals [33], as the latter would require expensive ground truth labels in order to train the network.

Our key contributions can be summarized as: (1) a class-conditional domain discriminator module (§ 3.3) which coupled with a gradient reversal layer promotes the learning of domain agnostic features; (2) an LSTM network (§ 3.2,3.5) to learn temporal domain agnostic features as complementary information; (3) state-of-the-art results on the four challenging domain generalization test sets (§ 4.2) with an accompanying visual analysis of the feature embedding (§ 4.4) and class activation maps (§ 4.6).

2. Related work

In what follows, we describe traditional, feature-based as well as modern, CNN-based approaches to FAS. We then elaborate on the few domain generalization works on FAS.

Traditional approaches. Before the advent of CNNs [26], typical approaches to face anti-spoofing combined the use of hand-crafted features with shallow classification techniques to detect differences in texture between the live and spoof images. The most characteristic examples of hand-crafted features include LBP [34], HoG [25], DoG [45], SIFT [38], and SURF [7]. In a similar vein, other traditional approaches employed heuristics to leverage 'liveliness' cues that are not present in a spoof attack. Examples of such heuristics are eyeblink [37] and lip motion [24]. Another ways to address face anti-spoofing are making use of temporal cues [4], different color spaces [5], image distortion analysis [48] or a transformation to the temporal domain [2] and Fourier spectrum [31], have been explored. In general, these traditional methods are either not applicable to all possible spoof attacks, *i.e*. print, replay, mask, or they fail to generalize to different datasets, since the learned features specialize to the 'trained' textures, which largely vary between datasets due to factors of variation like different backgrounds, lighting conditions, camera resolutions, spoof materials, *etc*.

CNN-based approaches. The impressive results achieved by applying CNNs to the tasks of image classification and object recognition [26, 19, 43, 44] motivated researchers to employ them to other computer vision tasks too. Face anti-spoofing is no exception. The obvious choice

[1]The term domain in this paper is used to refer to a dataset.

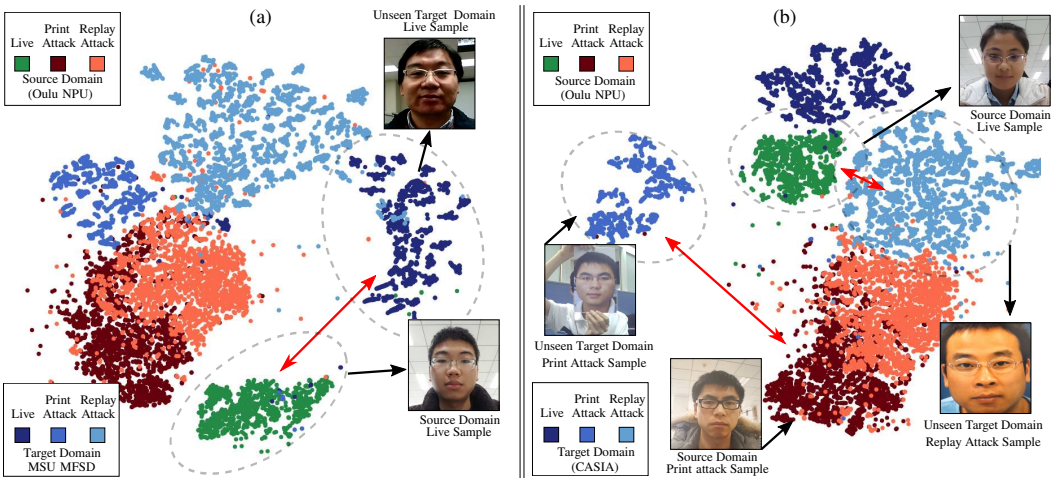


Figure 2. A t-SNE visualization of CNN features from a ResNet50 backbone trained on multiple source domains (i.e. FAS datasets) and tested on an unseen target domain. For better visualization we show only one source and one target domain in these plots. We can easily recognize the inherent domain shift problem in face anti-spoofing. That is, the live and spoof samples from the source and target domains are not properly aligned in the feature space, resulting in poor generalization of the learned feature representations on the target domain.

is to replace the hand-crafted features with features learned from generic CNNs - known for their great representational power when trained on large-scale datasets. Feng *et al*. [13] explored the use of multiple cues, such as image quality and motion cues. Xu *et al*. [49] incorporated video inputs and proposed an LSTM-CNN model to take advantage of the information from the extra frames. Dynamic textures were proposed in [41, 42] to extract different facial motions. Recently, Atoum *et al*. [1] introduced a multitasking-inspired approach that combines the estimation of texture and depth features for binary live/spoof classification, which was later extended by Liu *et al*. [33] to also include fusion with temporal supervision, *i.e*. rPPG signals. Finally, Joorabloo *et al*. [22] followed a different path and inversely decomposed a spoof face into a spoof noise and a live face using a GAN architecture, and consequently utilized the spoof noise for classification. Bresan et al. [9] explore depth, salience and illumination maps associated with a pre-trained CNN for FAS. They use combination of source domains (i.e. NUAA [45], Idiap Replay-Attack [10], CASIA-MFSD [52] dataset), different from ours, and thus their method is not directly comparable.

The aforementioned works, despite showing improved performance, partially attributed to the use of CNNs, still face open challenges when it comes to generalizing across domains (i.e. datasets). As mentioned, there is an inherent domain shift [30] between the different FAS datasets (*e.g*. Replay Attack [10] and CASIA-FASD [52]), which in turn leads to poor cross-testing results. In this paper, we go beyond current CNN-based approaches and explicitly tackle the domain shift problem in FAS without relying on supervision from extra cues, like depth or rPPG signals, that would require a significant annotation effort to acquire.

Domain generalization approaches. To tackle the domain shift problem across different datasets, domain adaptation [20, 14, 47, 15, 16] and generalization [23, 36, 50, 18, 17, 27, 35, 29] techniques have been used in computer vision. The goal in each case is to bridge the distribution gap between data from source and target domains in order to create domain agnostic feature representations that generalize to new domains. In this paper, we are mostly interested in domain generalization techniques, which have been largely unexploited in FAS, with the following exceptions. Li *et al*. [28] encouraged the learning of generalized feature representations by taking both spatial and temporal information into consideration and minimizing a cross-entropy loss together with a generalization loss. Tu *et al*. [46] proposed the use of Total Pairwise Confusion loss for CNN training in conjunction with a Fast Domain Adaptation component into the CNN model to account for domain changes. Shao *et al*. [40] combined the learning a generalized feature space that is shared by multiple discriminative source domains with dual-force triplet mining constraint to improve the discriminability of the learned feature space. In general, compared to the aforementioned works our framework offers better integration to multiple domains, and, as will be shown in Sec. 4, achieves significantly improved results on four public datasets.

3. Proposed Approach

3.1. The domain shift problem in face anti-spoofing

Our main goal is to learn generalized feature representations in order to address the domain shift problem that in-

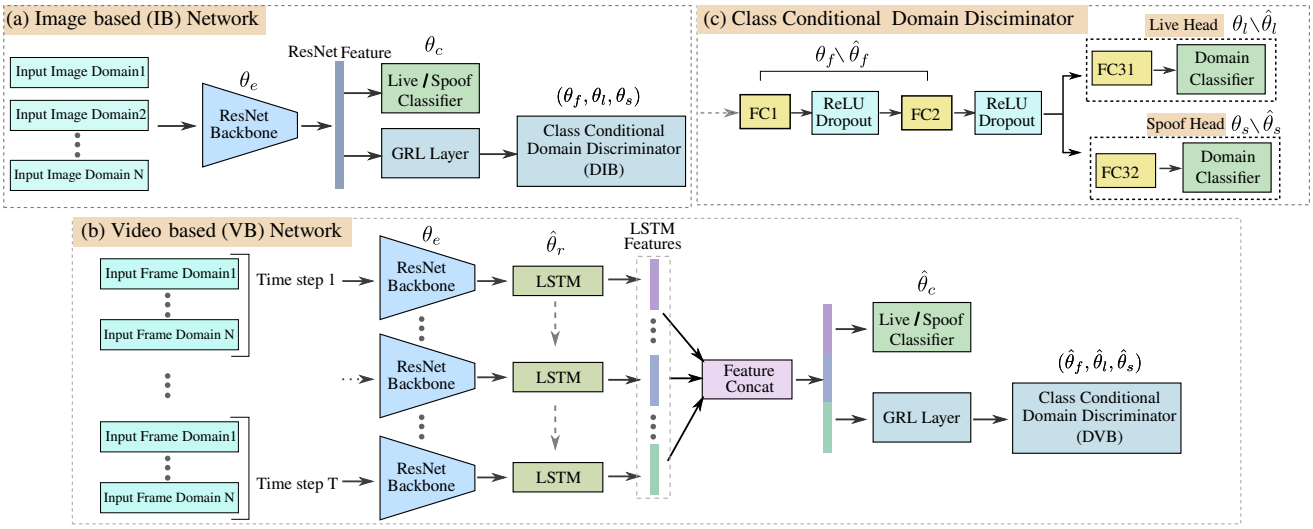


Figure 3. Overview of the different components of the proposed approach. See Section 3 for more details.

herently exists among FAS datasets. That is, the distribution dissimilarities between live and spoof samples that belong to multiple source and unseen target domains. To illustrate this problem, we use t-SNE plots (Fig. 2) generated from the CNN features of a ResNet50 [21] backbone trained on multiple source domains (i.e. FAS datasets) for live/spoof classification, and tested on an unseen target domain. As can be seen in Fig. 2 (a), the CNN features of the live samples from the unseen target domain are far away from the live samples of the source domain in the feature space. Similarly in Fig. 2 (b), we can see that the print attack features from the target domain are far apart from the source domain's print attacks, and the target domain's replay attack features are shifted towards the live samples of the source domain. It is quiet evident from these illustrations that even deep neural networks, like ResNet models, are not sufficient on their own to tackle the problem. This calls for dedicated mechanisms that can leverage the common attributes shared across multiple source domains to learn more generic feature representations. The term common attributes is used here to refer to the common intrinsic properties of the print and replay attacks across multiple domains. For example, although these attacks might have been generated using different spoofing mediums (i.e. different printers or video capturing devices), or under different environmental conditions (e.g. illumination, background scene), they are inherently based on paper materials or display screens. Thus, by leveraging these common attributes one could expect that better feature representations can be learned from the shared and discriminative information across multiple source domains, that is robust for live/spoofing classification and at the same time domain agnostic. We expect such representations to demonstrate better generalization on unseen target domains.

3.2. System overview

To tackle the aforementioned problem, we propose a novel framework which learns both image- and video-based domain agnostic feature representations (see Fig. 3). More specifically, a ResNet backbone (encoder) is trained to minimize the live/spoof classification loss, while at the same time it competes against a class-conditional domain discriminator (§3.3) coupled with a gradient reversal layer to maximize the domain classification loss of live and spoof samples respectively. During the training process the encoder gradually learns the shared and discriminative feature representations. A system overview is given in Fig. 3.

You can observe two variations (see Fig. 3). First, an image based (IB) network that follows an image-level training, in which a training example consists of an image and its associated ground-truth label (either "live" or "spoof"). This is to demonstrate the scenario where only a single image is given as input, and the system has to decide if this is a spoof attack or not. However, FAS can also be a video classification problem, i.e. we expect the final output to be a live/spoof label for an input video sample. Thus, a CNN trained following an image-level protocol might fail if we process the results on a frame-by-frame basis, as the video itself usually contains richer information. For such instances, we want the network to learn strong temporal features which are complementary to the spatial representation learned by the IB network. Based on this idea, we also propose a video-based (VB) network which is trained along-side the IB network, following an alternating training scheme [33]. This VB network uses the same ResNet backbone, i.e. model parameters of the ResNet backbone are shared between the IB and VB networks. Unlike the IB network, the VB network inputs video sequence and

processes these through multiple long-short term memory (LSTM) units and outputs a single class label for each input video sequence.

3.3. Class-conditional domain discriminator

In Fig. 3 (c), we show the network architecture of our proposed class-conditional domain discriminator (CCDD). CCDD consists of two fully connected layers, FC1 and FC2, followed by a *live* and a *spoof* head. FC1 and FC2 layers are followed by a ReLU and a dropout layer. During training, an SGD mini-batch that consists of live and spoof training examples is processed through the FC1 and FC2 layers. Consequently, the outputs of the FC2 layer are first split into "live" and "spoof" batches, and then, they are passed as input to their respective heads. The *live* and *spoof* heads have the same layer configuration, i.e. each consists of a single linear transformation layer followed by a domain classifier. They output two score vectors $\mathbf{s}_l$ and $\mathbf{s}_f$ having D scores, *i.e*. the softmax probability scores for each domain. Note that, we use the same network architecture for the image- and video-based CCDD (DIB and DVB in Fig.3 (a) & (b)).

The proposed CCDD coupled with the gradient reversal layer imposes the desired conditional invariance property on the learned feature representations. The conditional invariance is realized by the class-conditional losses (see below), which consider the source domain label information only and aim to make the representation in each class indistinguishable across domains. We present a t-SNE visualization (§4.4) to demonstrate that the proposed CCDD learns to correctly align the live and spoof features of the target domain with the features of source domains. Besides, we present quantitative experimental results to attest the effectiveness of the CCDD. A more detailed network design is provided in the supplementary material.

3.4. Gradient reversal layer

The gradient reversal layer (GRL) [15] was originally proposed for unsupervised domain adaptation. Instead, we couple CCDD with GRL in order to learn domain agnostic features from multiple source domains for FAS. In particular, we use two GRL layers, one in the image-based and another one in the video-based network (Fig.3). What GRL essentially does, is to reverse the gradient by multiplying it by a negative scalar (i.e. the adaptation factor λ_{GRL}) during the backward propagation. During the forward propagation, it leaves the input unchanged, *i.e*. it acts as an identity transform. By doing so, it essentially reverses the objective of its subsequent layers, i.e. CCDD in our case. What this practically means, is that the backbone network is now tasked with the extra objective of generating live and spoof feature representations that are indistinguishable across multiple source domains.

3.5. Optimization cost

First, we specify the energy function used to optimize the IB network (Fig.3 (a)). Consider the following notations: θ_f, θ_l and θ_s be the model parameters of the common layers (i.e. FC1 & FC2), *live* and *spoof* heads of the DIB respectively; θ_e and θ_c be the model parameters of the encoder (i.e. the ResNet backbone) and the label classifier (i.e. the live/spoof classifier); L_l and L_s be the domain classification losses (*i.e*. multinomial) for the *live* and *spoof* heads that penalize for incorrect domain label prediction separately for the "live" and "spoof" training examples; L_c be the label classification (e.g. multinomial) loss that penalizes for incorrect class label (i.e. "live" or "spoof") prediction; i denotes the index for a training example and F be the number of training examples, *i.e*. $i = \{1, 2, \ldots, F\}$; b_i be a binary variable denoting the class label of the i-th example, *i.e*. $b_i = 0$ indicates that the example is live and $b_i = 1$ that it is a spoof. During the IB network training, the encoder's model parameters θ_e learn to minimize the discrepancy in the class conditional distribution [32] across different domains. This is done by maximizing the domain classification losses of the *live* and *spoof* heads of the DIB. In other words, it tries to make the feature distributions (belonging to a class $c \in C$) maximally similar across different domains. At the same time, the *live* and *spoof* heads seek parameters θ_l and θ_s which minimize the class conditional domain classification losses. This yields as energy function for our IB network:

$$E(\theta_e, \theta_c, \theta_f, \theta_l, \theta_s) = \sum_{i=1\ldots F} L_c^i(\theta_e, \theta_c) + \lambda_{IB}\left(\sum_{\substack{i=1\ldots F \\ b=0}} L_l^i(\theta_e, \theta_f, \theta_l) + \sum_{\substack{i=1\ldots F \\ b=1}} L_s^i(\theta_e, \theta_f, \theta_s)\right) \tag{1}$$

Now, we specify the energy function used to optimize the VB network (Fig.3 (b)). Let: $\hat{\theta}_r$ be the model parameters of the LSTM network; $\hat{\theta}_f$, $\hat{\theta}_l$ and $\hat{\theta}_s$ be the model parameters of the common layers (i.e. FC1 & FC2), *live* and *spoof* heads of the video-based class-conditional domain discriminator respectively; $\hat{\theta}_c$ be the model parameters of the LSTM's label classifier (i.e. the live/spoof classifier). In a similar fashion, during the VB network training the encoder's and LSTM's model parameters (i.e. θ_e and $\hat{\theta}_r$) learn to minimize the discrepancy in the class conditional distribution across different domains by maximizing the domain classification losses of the *live* and *spoof* heads of the DVB. At the same time, the *live* and *spoof* heads seek parameters $\hat{\theta}_l$ and $\hat{\theta}_s$ which minimize the class conditional domain classification losses. This yields as energy function for our VB network:

Table 1. Comparison to state-of-the-art FAS methods on four domain generalization test sets.

Method	O&C&I→M		O&M&I→C		O&C&M→I		I&C&M→O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MS LBP [34]	29.76	78.50	54.28	44.98	50.30	51.64	50.29	49.31
Binary CNN [51]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
IDA [48]	66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
Color Texture [6]	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
LBPTOP [12]	36.90	70.80	42.60	61.05	49.45	49.54	53.15	44.09
Auxiliary(Depth Only) [33]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
Auxiliary(All) [33]	-	-	28.4	-	27.6	-	-	-
Ours	**15.42**	**91.13**	**17.41**	**90.12**	**15.87**	**91.72**	**14.72**	**93.08**

Table 2. Comparison to state-of-the-art domain generalization FAS methods on four domain generalization test sets.

Method	O&C&I→M		O&M&I→C		O&C&M→I		I&C&M→O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MMD-AAE [29]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
MADDG [40]	17.69	88.06	24.5	84.51	22.19	84.99	27.98	80.02
Ours	**15.42**	**91.13**	**17.41**	**90.12**	**15.87**	**91.72**	**14.72**	**93.08**

$$E(\theta_e,\hat{\theta}_r,\hat{\theta}_c,\hat{\theta}_f,\hat{\theta}_l,\hat{\theta}_s) = \sum_{i=1...F} \hat{L}_c^i(\theta_e,\hat{\theta}_r,\hat{\theta}_c) + \lambda_{VB}\left(\sum_{\substack{i=1...F\\ b=0}} \hat{L}_l^i(\theta_e,\hat{\theta}_r,\hat{\theta}_f,\hat{\theta}_l) + \sum_{\substack{i=1...F\\ b=1}} \hat{L}_s^i(\theta_e,\hat{\theta}_r,\hat{\theta}_f,\hat{\theta}_s)\right) \quad (2)$$

$\hat{L}_c^i$, $\hat{L}_l^i$ and $\hat{L}_s^i$ are the live/spoof classification loss and the domain classification losses (for the live and spoof heads) for VB network. λ_{IB} and λ_{VB} are the scalar parameters weighting the relative importance of the two loss terms in Eq. 1 and Eq. 2 respectively. Note that, the encoder's model parameters θ_e are shared across the image- and video-based networks.

4. Experiments

4.1. Experimental setting

Datasets. We evaluate our method on four publicly available FAS datasets: Oulu-NPU [8] (O for short), CASIA-MFSD [52] (C for short), Idiap Replay-Attack [10] (I for short), and MSU-MFSD [48] (M for short).

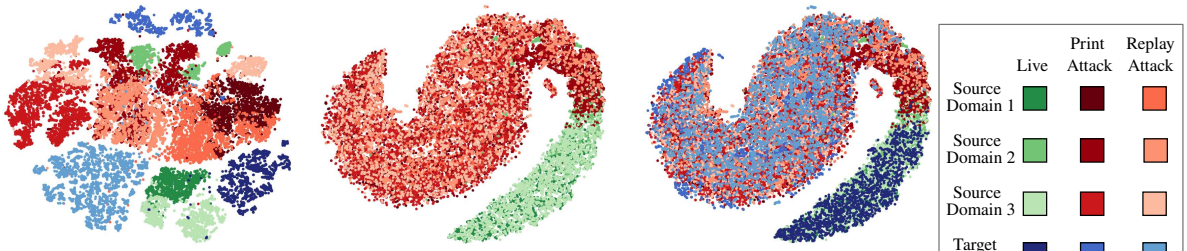
Training and evaluation. We consider a dataset to be one domain in our experiments. Our model learns domain generalized representations from three out of four datasets, as in [40]. In particular, we randomly select three datasets as the source domains, and the remaining unseen domain, which is not accessed during training, is kept for evaluation only. Half Total Error Rate (HTER) [3] and Area Under Curve (AUC) are used as the evaluation metrics in our experiments.

Implementation details. We use ResNet-50 [21] as our backbone network. The dimension of the input image is 224×224. During training, we use SGD optimizer, and follow an alternative training approach [33] to train both our IB and VB networks (Fig. 3). We use a constant learning rate of 0.0003, momentum 0.9 and weight decay 0.00001. The mini-batch size for the IB network is 48, i.e. 16 training images from each of the three domains. For the VB network, the mini-batch size is 6, i.e. 2 training video sequences from each of the three source domains, and the LSTM sequence length is 8. The LSTM's input dimension is 2048, while the hidden layer dimension is 256. We use a constant GRL adaptation factor ($\lambda_{GRL} = -0.2$) [15], and set the λ_{IB} and λ_{VB} to 1. Additional experimental details are presented in the supplementary material.

4.2. Comparison to the state-of-the-art

In Table 1, we compare our full model against state-of-the-art FAS methods. Our proposed method outperforms [34, 51, 48, 6, 12, 33] on all the four domain generalization test sets. The significantly better performance mostly lies in the ability to learn rich generalizable features, which adapt well to the unseen target domain (see Fig.4). Note that, these FAS methods do not explicitly address the domain shift problem, and thus naturally fail to generalize well on unseen target domains. In contrast, our proposed method explicitly learns a generalizable representation by leveraging the available information (live and spoof examples with ground truth labels) from multiple source domains. In particular, it learns to map all the live and spoof samples (from multiple source domains) to a common feature space where the live and spoof features are far apart, while being domain invariant at the same time.

In addition, we compare against the state-of-the-art domain generalization FAS method [40] and also compare to the related state-of-the-art method in domain generalization for the face anti-spoofing task: MMD-AAE [29] as in [40]. These methods explicitly address the domain shift problem. Table 2 shows this comparison, where our method consistently achieves much better performance. We conclude that the proposed method can overcome the distribution dissimilarities in the feature space more effectively. Moreover, [40] is relatively expensive and not end-to-end trainable, in con-



(a) ResNet 50 Visualizing 3 source and 1 target domains

(b) Proposed Model Visualizing only 3 source domains

(c) Proposed Model Visualizing 3 source and 1 target domains

Figure 4. A t-SNE plot of the CNN features coming from ResNet (a) vs our full model (b,c), both trained on three source domains and tested on an unseen target domain (best viewed in color). Note that, the live features of source and target domains are far apart (a); similar trend can be noticed for the spoof features of source and target domain, but our model learns to group together all live and spoof features (from multiple source domains) into two different clusters (b), thus improving the classification accuracy. Importantly, the learned representations generalize well on the target domain (c).

Table 3. An ablation study of the different components in the proposed FAS architecture on four domain generalization test sets.

ResNet	DIB	LSTM	DVB	O&C&I→M		O&M&I→C		O&C&M→I		I&C&M→O	
				HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
✓				21.66	89.64	25.92	82.16	20.12	90.1	18.81	89.53
✓	✓			18.33	90.58	21.29	85.82	17.63	86.3	17.05	90.01
✓		✓		17.92	90.27	19.26	87.85	18.0	89.78	16.42	90.82
✓		✓	✓	18.33	88.25	21.11	88.22	18.25	85.61	17.05	91.09
✓	✓	✓		**14.58**	**92.58**	18.7	89.35	**15.13**	**95.76**	14.86	93.00
✓	✓	✓	✓	15.42	91.13	**17.41**	**90.12**	15.87	91.87	**14.72**	**93.08**

trast to our method.

4.3. Ablation study on model components

So far, we have shown results with our *full model* that contains all the different components, i.e. ResNet backbone (ResNet), image-level domain discriminator (DIB), LSTM module (LSTM), and video-level domain discriminator (DVB). In what follows, we present a detailed ablation study when using different combinations of these components. The experimental results on all four domain generalization test sets are summarized in Table 3. When we mention DIB or DVB in Table 3, it automatically includes the associated GRL layer.

To demonstrate the applicability of the proposed model components, we first setup our own baseline for the ablation study. The baseline is comprised of a ResNet-50 backbone and a live/spoof classifier which is trained on the four different domain generalization training sets. Our baseline itself exhibits some desirable performance. In the supplementary material, we report experiments with a lighter ResNet backbone. When adding DIB on top of the ResNet backbone, the results are consistently improved on all four test sets. Additionally adding LSTM, the results are again improved significantly. Finally, our full model boosts the results further. Combining ResNet and LSTM, provides slightly better results on three test setups compared to the model using ResNet and DIB. However, adding DVB to the model with ResNet and LSTM does not bring any further improvements. However, when DVB is jointly trained with ResNet, DIB and LSTM, i.e. our full model, improves over the ResNet baseline. This observation verifies that by exploiting both spatial (DIB or image-based) and temporal (DVB or video-based) domain-agnostic features our proposed model can achieve the best results on the two most challenging domain generalization test sets (O&M&I→C and I&C&M→O).

4.4. Visualization of the learned CNN features

Fig. 4 depicts t-SNE plots of the CNN activations (*i.e.* features) coming from our ResNet baseline vs our full model. Both networks were trained on 3 source domains (*i.e.* Oulu-NPU, CASIA-MFSD and MSU-MFSD) and tested on a target domain (*i.e.* Idiap replay-attack). Note that, the plots in (b) and (c) are generated using the same trained model, i.e. our full model, and the same set of live and spoof samples. For the sake of better visualization, however, we have deactivated the visualization of the target domain in (b). As can be seen in (b), our model learns more discriminative features for live and spoof images. What is more interesting is that the representation learned by our model aligns well with unseen target domain's live and spoof features, as can be seen by activating the target domain visualization in (c). In contrast, the ResNet learnt representation shows relatively weaker generalization ability on the target domain, as shown in (a). In the latter case, the live, print- and replay-attack features from

multiple source domains are far apart in the feature space, whereas our model learns to minimize this inter-domain distances between live and spoof features, as shown in (b, c). From these visualizations, we can conclude that our network generalizes well on the target domain. Particularly observe in (c) how the target domain live and spoof features are properly aligned with the live and spoof features of the source domains in (b).

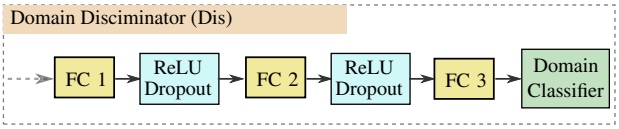


Figure 5. Architectural components of our default Domain Discriminator network (Dis).

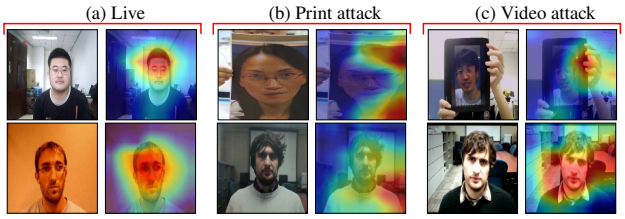


Figure 6. Activation map visualization of the proposed network. For each column (a), (b) and (c), the original input images and its associated network class activation maps are shown.

Table 4. Performance comparison of different domain discriminators on three domain generalization test sets.

ResNet	Dis	DIB	S&O&I&R →C HTER(%)	S&O&C&R →I HTER(%)	S&C&I&R →O ACER(%)
✓			17.5	20.6	10.27
✓	✓		15.3	17.7	8.75
✓	✓	✓	15.1	17.0	23.4
✓		✓	**14.0**	**14.7**	**8.05**

4.5. Impact of different domain discriminators

We conduct experiments to analyze the effect of using different domain discriminators on the FAS performance. We consider two domain discriminator architectures: the proposed DIB (Fig. 3), and the default domain classifier (Fig. 5) originally proposed by [15] for unsupervised domain adaptation (Dis in Table 4). Note that, for the experiments in this section we used – only for training purposes – two more datasets, i.e. SiW (S for short) [33] and Idiap replay-mobile (R for short) [11]. Following [33], when testing on Oulu-NPU dataset, we use the ACER metric. From Table 4, it can be seen that our ResNet-DIB gives the best performance. When ResNet-Dis is used, the performance degrades slightly. Even combining Dis with DIB degrades the performance heavily on Oulu-NPU. From these experiments, we observe that learning feature representations from multiple source domains conditioned on class labels (*i.e.* live and spoof) can provide discriminative and domain agnostic features, while conditioning them on domain labels only may not correctly align the live and spoof features, resulting poor classification accuracy. As the proposed DIB has access to both class (live and spoof) and domain labels, in contrast to Dis, it is able to learn better representations by correctly grouping live features from multiple source domain into one cluster and spoof features into another (see Fig. 4).

4.6. Class activation map visualization

In this section, we provide a visual analysis of the class activation maps to get an intuition about the decisions the network makes when making a particular prediction. For this visualization, we use the Grad-CAM [39] technique. In Fig. 6 we show the class activation maps for the live, print and replay attack test samples. Some interesting observations can be made. The network gives more importance to the facial regions for detecting a "live" class (see Fig. 6 (a)) which is intuitive as most of the information about a live face comes from the facial region. For example, the texture of a live skin, the eye blinking, head motion etc. On the other hand, for print attacks the network pays more attention to the surface of the paper (on which the face image is printed) (Fig. 6 (b)). For video replay attacks, if strong features like "a hand in the background" and "a tablet screen" are present then the network takes decision from these salient information (Fig. 6 (c) top). In the absence of such strong features, it tries to see both the facial region and the background (Fig. 6 (c) bottom).

5. Conclusion

In this paper, we addressed an inherent problem in face anti-spoofing, i.e. the large variability in factors such as the different backgrounds, lighting conditions, camera resolutions, spoof materials, etc., makes feature representations learned by CNNs for this task too domain-dependent, leading to decreased performance when testing on unseen domains. We propose a solution based on generalizable feature learning that naturally fits this 'domain shift' problem in both image-based and video-based face anti-spoofing. We provide extensive experimentation on multiple aspects of our approach, and among others, we demonstrate state-of-the-art performance across different test sets, we illustrate the qualitative improvement of the learned feature representations w.r.t. generalization, and visualize through the class activation maps the network's attention when making predictions. For future work, we would like to use multi-modal inputs and apply domain agnostic multi-modal feature learning to further improve the classification accuracy.

References

[1] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328. IEEE, 2017. 3

[2] Wei Bao, Hong Li, Nan Li, and Wei Jiang. A liveness detection method for face recognition based on optical flow field. In *2009 International Conference on Image Analysis and Signal Processing*, pages 233–236. IEEE, 2009. 2

[3] Samy Bengio and Johnny Mariéthoz. A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, number CONF, 2004. 6

[4] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Face anti-spoofing via motion magnification and multifeature videolet aggregation. Technical report, 2014. 2

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*, pages 2636–2640. IEEE, 2015. 2

[6] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016. 6

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

[8] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. 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)*, pages 612–618. IEEE, 2017. 1, 6

[9] Rodrigo Bresan, Allan Pinto, Anderson Rocha, Carlos Beluzo, and Tiago Carvalho. Facespoof buster: a presentation attack detector based on intrinsic image properties and deep learning. *arXiv preprint arXiv:1902.02845*, 2019. 3

[10] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012. 1, 2, 3, 6

[11] Artur Costa-Pazo, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sebastien Marcel. The replay-mobile face presentation-attack database. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2016. 8

[12] Tiago de Freitas Pereira, Jukka Komulainen, André Anjos, José Mario De Martino, Abdenour Hadid, Matti Pietikäinen, and Sébastien Marcel. Face liveness detection using dynamic texture. *EURASIP Journal on Image and Video Processing*, 2014(1):2, 2014. 6

[13] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. 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. 3

[14] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013. 3

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 2, 3, 5, 6, 8

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 3

[17] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2017. 3

[18] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 3

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[20] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pages 999–1006. IEEE, 2011. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[22] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face despoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 290–306, 2018. 2, 3

[23] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 3

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

[25] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2013. 2

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017. 3

[28] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018. 3

[29] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 3, 6

[30] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. 2, 3

[31] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain. Live face detection based on the analysis of fourier spectra. In *Biometric Technology for Human Identification*, volume 5404, pages 296–304. International Society for Optics and Photonics, 2004. 2

[32] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 5

[33] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018. 2, 3, 4, 6, 8

[34] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011. 2, 6

[35] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017. 3

[36] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013. 3

[37] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[38] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283, 2016. 2

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[40] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. 2, 3, 6

[41] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 748–755. IEEE, 2017. 3

[42] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 14(4):923–938, 2018. 3

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[45] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517. Springer, 2010. 2, 3

[46] Xiaoguang Tu, Jian Zhao, Mei Xie, Guodong Du, Hengsheng Zhang, Jianshu Li, Zheng Ma, and Jiashi Feng. Learning generalizable and identity-discriminative representations for face anti-spoofing. *arXiv preprint arXiv:1901.05602*, 2019. 3

[47] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 3

[48] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 1, 2, 6

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

[50] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. 3

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

[52] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *ICB*, pages 26–31. IEEE, 2012. 1, 3, 6