Learning Facial Liveness Representation for Domain Generalized Face Anti-Spoofing

Zih-Ching Chen*, Lin-Hsi Tsao*, Chin-Lun Fu*, Yu-Chiang Frank Wang

Abstract

Face anti-spoofing aims at distinguishing face spoof attacks from the authentic ones, which is typically approached by learning proper models for performing the associated classification task. In practice, one would expect such models to be generalized to FAS in different image domains. Moreover, it is not practical to assume that the type of spoof attacks would be known in advance. In this paper, we propose a deep learning model for addressing the aforementioned domain-generalized face anti-spoofing task. In particular, our proposed network is able to disentangle facial liveness representation from the irrelevant ones (i.e., facial content and image domain features). The resulting liveness representation exhibits sufficient domain invariant properties, and thus it can be applied for performing domain-generalized FAS. In our experiments, we conduct experiments on five benchmark datasets with various settings, and we verify that our model performs favorably against state-of-the-art approaches in identifying novel types of spoof attacks in unseen image domains.

Introduction

Face recognition technology has been widely applied in many interactive intelligent systems such as automated teller machines (ATMs), mobile payments, and entrance guard systems, due to their convenience and remarkable accuracy. However, face recognition systems are still vulnerable to presentation attacks (PAs) ranging from print attacks, video replay attacks, and 3D facial mask attacks, etc. Therefore, face anti-spoofing (FAS) plays a crucial role in securing the robustness of face recognition systems.

Over the past few years, different FAS methods have been proposed by researchers. Assuming inherent disparities between live and spoof faces, studies have handled this problem from the perspective of detecting texture in color space (Boulkenafet, Komulainen, and Hadid 2015; Chingovska, Anjos, and Marcel 2012), image distortion (Wen, Han, and Jain 2015), temporal variation (Shao, Lan, and Yuen 2017), or deep semantic features (Yang, Lei, and Li 2014; Patel, Han, and Jain 2016a). Although promising results have been obtained under the intra-database testing scenarios by the existing FAS methods, a major open challenge in FAS is the domain shift problem, in which face recognition models are trained on one or multiple datasets (i.e., source domains), while it is expected to perform FAS in unseen datasets (i.e., target domains). Without proper handling of such domain shifts, the model performance would degrade dramatically.

To address the above unseen domain shift problem mentioned above, studies have exploited auxiliary information such as face depth (Liu, Jourabloo, and Liu 2018), which is scenario-invariant for distinguishing live and spoof faces. However, these approaches still have their limitations since they highly depend on the accuracy of estimated auxiliary information. Recently, researchers start to improve the robustness of FAS from the perspective of domain generalization, which aims to learn a generalized feature space by aligning the distributions among multiple source domains. Shao (Shao et al. 2019) assume that the extracted features of unseen faces can be mapped to a shared feature space so that the model can generalize to the unseen domains. They proposed a multi-adversarial deep domain generalization method to find two generalized feature spaces for the real images and the fake ones, respectively. Holding a different point of view, Jia (Jia et al. 2020) assume that it is hard to map the features of fake faces from different domains to one single generalized feature space since the attack types and collecting ways can be very different. Therefore, they proposed a single-side domain generalization method pulling all the real faces to one generalized feature space while separating the fake ones of different domains. However, the generalization ability of the mentioned approaches might be limited. This is because the existing methods typically do not distinguish between domain-independent facial liveness representations and the domain-dependent ones which are irrelevant to FAS (Wang et al. 2020).

In light of the above issue, researchers handled FAS from the aspect of feature disentanglement. A multi-domain disentangled representation learning method is proposed by (Wang et al. 2020), aiming to obtain more discriminative presentation attacks detection (PAD) features by disentangling domain-invariant representations from an image. While (Wang et al. 2020) have disentangled the domain-independent FAS cues from the domain-dependent representations, their approach might not be able to generalize to real-world scenarios, in which novel (i.e., unseen) spoof

*These authors contributed equally.

attacks might be presented. Although existing works focus on disentangling domain-relevant features from other features, their extracted liveness features still contain facial content information, which is irrelevant to liveness information. Thus, the generalization ability of their methods to handle unseen spoof attacks is still limited.

In this paper, we address the domain-generalized FAS problem by learning domain-invariant facial liveness representation. We not only aim to handle FAS in unseen data domains but unseen spoof attacks can also be detected, which makes our proposed model more practical. As detailed later, we present a representation disentanglement framework, which is designed to extract facial liveness, content, and image domain representations. More specifically, the liveness representation describes information for liveness detection. On the other hand, the latter two representations, i.e., the facial content and image domain representations are viewed as liveness-invariant features. The disentanglement of such features from the liveness features allows our model to better perform FAS in unseen domains with novel spoof attacks. The contributions of this work can be highlighted below:

- We propose a representation disentanglement network for domain-generalized face liveness detection, which is able to recognize novel spoof attacks in unseen domains/datasets during inference.
- Our proposed network is designed to extract facial liveness, content, and image domain representations. While the liveness representation would be utilized for FAS, the latter two are liveness-invariant.
- We conduct experiments on multiple FAS datasets in various settings. We confirm that our proposed method performs favorably against state-of-the-art approaches in detecting novel spoof attacks in unseen image domains.

Related Work

Face Anti-Spoofing

The problem of FAS is typically approached by the following strategies. Utilizing handcrafted feature descriptors, previous methods including LBP (Boulkenafet, Komulainen, and Hadid 2015; de Freitas Pereira et al. 2012, 2013; Määttä, Hadid, and Pietikäinen 2011), HoG (Komulainen, Hadid, and Pietikäinen 2013; Yang et al. 2013), SIFT (Patel, Han, and Jain 2016b), and SURV (Boulkenafet, Komulainen, and Hadid 2016) have been applied for solving FAS using the associated features, followed by a trained binary classifier for predicting the authenticity of the input face image. With the development of deep learning approaches, various methods apply convolutional neural networks (CNN) (Krizhevsky, Sutskever, and Hinton 2012) to extract more discriminative features, achieving significant improvements. Feng (Feng et al. 2016) regarded FAS as a binary classification task via a CNN architecture. Additional auxiliary supervision such as depth map, reflection map, and r-ppg signals are also considered for improving the FAS classification performances. Recently, Zhang (Zhang et al. 2020a) and Liu (Liu, Stehouwer, and Liu 2020) extracted liveness features from the perspective of disentanglement. Although such techniques have shown promising performances, it is not clear how their extracted liveness features can be generalized across image domains/datasets, or to handle novel spoof attacks.

Domain Generalization

Domain generalization aims at leveraging information learned from source domains to target domains that are unseen during training. To address FAS across image domains, Shao (Shao et al. 2019) proposed a multi-adversarial discriminative deep domain generalization (MADDG) framework. With a dual-force triplet-mining constraint, their proposed architecture learns a generalized feature space for real and fake faces. Extended from MADDG, Jia (Jia et al. 2020) designed a single side triplet-mining algorithm, which performs FAS while grouping spoof types across domains for the purpose of domain generalization. On the other hand, Liu (Liu et al. 2021) proposed a Dual Reweighting Domain Generalization (DRDG) framework which reweights the samples from different domains during training, aiming to separate real and fake images disregard their image domains. While the above works have shown improved generalization ability on unseen domains for FAS, it is not clear whether they can handle novel spoof attacks, and thus their usage for practical FAS tasks might still be limited.

Representation Disentanglement

Representation disentanglement focuses on extracted particular feature representations, which would describe the information of interest, typically benefiting the downstream learning tasks. For FAS, (Wang et al. 2020) proposed to jointly perform disentangled representation learning and multi-domain feature learning for cross-domain face PAD. (Zhang et al. 2020a) chose to disentangle the liveness features from the feature representation of the input image, with both low-level and high-level auxiliary supervision to improve the generalization abilities. Nevertheless, these methods did not take the scenario that the target domain may not be presented during training, and the type of spoof attacks might not be novel to the existing ones. This is the reason why, as we detail in the following section, we propose to learn domain-invariant facial liveness representation for FAS. Moreover, since liveness-irrelevant information such as facial content and image domains are disentangled from the derived liveness representation, domain-generalized FAS can be achieved.

Proposed Method

Problem Definition and Annotations

For the sake of completeness, we first define the setting and notations considered in this paper. During training, we have face images in S different source domains denoted as $X = \{X_1, X_2, ..., X_S\}$ and the corresponding binary real/fake labels denoted as $Y = \{Y_1, Y_2, ..., Y_S\}$. For the ith source domain, we have N_i images, i.e., $X_i = \{x_{i,j}\}_{j=1}^{N_i}$, and the associated labels $Y_i = \{y_{i,j}\}_{j=1}^{N_i}$. As for the inference stage, liveness of facial images in a disjoint and unseen target domain (with seen attacks and unseen attacks) will be determined accordingly.

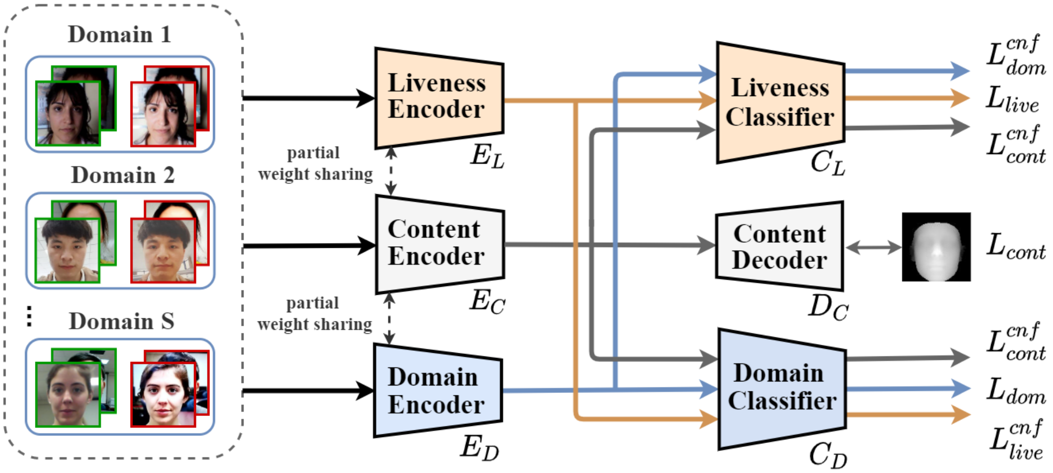

Figure 1: Overview of our proposed network architecture. Our network aims to extract liveness, facial content, and image domain representations from facial input images. This is realized by the learning of liveness encoder E_L, content encoder E_C, and domain encoder E_D, respectively. To further ensure the disentanglement of liveness-irrelevant information, liveness classifier C_L, content decoder D_C, and domain classifier C_D are jointly deployed. Once the training is complete, one can apply E_L and C_L for domain-generalized FAS.

As shown in Fig. 1, our proposed network architecture has three encoders for handling the facial input images: liveness encoder E_L, content encoder E_C, and domain encoder E_D. The liveness encoder E_L is designed to extract liveness representation, followed by a liveness classifier C_L for performing FAS prediction. The content encoder E_C extracts the facial content representation from the input, while the subsequent decoder D_C is deployed for reconstruction guarantees. As for the encoder E_D, it extracts the domain representations from the training input images, so that the domain classifier C_D would classify the image domain accordingly. Once the joint learning of the above network modules is complete, one can simply take the liveness encoder E_L and the liveness classifier C_L for inference.

Learning Liveness-Irrelevant Representation

In order to address domain-generalized FAS, we propose to extract facial content and image domain features from the liveness representation. Since the disentangled content and image domain features are not utilized for FAS, they can be viewed as liveness-irrelevant representations of face images.

As depicted in Fig. 1, the above two types of features are extracted by the content encoder E_C and domain encoder E_D. For the content encoder, it is expected to retrieve facial content information, which is not regarding the authenticity of the face input image. In our work, we specifically consider content information described by PRNet (Feng et al. 2018), which is known for describing 3D face information by observing a face image. The idea is that spoof image or not, the facial input image is expected to contain facial contour and landmark information, which suggest the recovery of the 3D shape. Thus, given the jth training image from source domain i, the encoded content feature of $E_C(x_{i,j})$ would serve as the input to the content decoder D_C, which is designed to recover the feature produced by a pre-trained PRNet $\Phi(\cdot)$. In other words, we calculate the following *content loss* L_{cont} for updating both E_C and D_C:

$$L_{cont} = \sum_{i=1}^{S}\sum_{j=1}^{N} \|D_C(E_C(x_{i,j})) - \Phi(x_{i,j})\|_2^2. \tag{1}$$

While the calculation of (1) ensures E_C and D_C for extracting and recovering facial content information, we need additional supervision to ensure the derived content features $E_C(x_{i,j})$ does not contain either liveness or image domain information. As a results, with the deployment of the liveness classifier C_L and domain classifier C_D, we choose to calculate the following *content confusion loss* L_{cont}^{cnf}:

$$\begin{aligned} L_{cont}^{cnf} = \sum_{i=1}^{S}\sum_{j=1}^{N}(\|C_L(E_C(x_{i,j})) - \frac{1}{2}\|_2^2 \\ +\|C_D(E_C(x_{i,j})) - \frac{1}{S}\|_2^2). \end{aligned} \tag{2}$$

The first and second terms in (2) are to confuse the liveness and domain classifiers, respectively. It is worth noting that, C_L predicts the binary liveness label, and C_D outputs the domain label (out of S). With the observation of this content confusion loss, the training of our proposed framework would further ensure the content encoder E_C to produce liveness and domain-irrelevant features.

As for the learning of image domain features, the modules of domain encoder E_C and domain classifier C_D are deployed for achieving this goal. Following the above example, assume that we have the jth training image from source domain i, the encoded domain feature $E_D(x_{i,j})$ is expected to describe illumination, image quality, etc. information. Thus, the subsequent domain classifier C_D is designed to recognize the image domain i by observing such domain features. In our work, we calculate the *domain loss* L_{dom} for updating both E_D and C_D as follows:

$$L_{dom} = -\sum_{i=1}^{S}\sum_{j=1}^{N} m_i * log(C_D(E_D(x_{i,j}))). \tag{3}$$

In the above equation, m_i denotes the ground truth one-hot vector representing the domain label.

Similar to the design of liveness-irrelevant facial content features, we now discuss how we further ensure that the learning of E_D and C_D would not contain liveness information. With the deployment of the liveness classifier C_L in Fig. 1, we have C_L take the encoded domain features $E_D(x_{i,j})$. To enforce the disentanglement of liveness-relevant information, the liveness classifier C_L is not expected to perform FAS on $E_D(x_{i,j})$. Therefore, we calculate the *domain confusion loss* L_{dom}^{cnf}, which is defined below:

$$L_{dom}^{cnf} = \sum_{i=1}^{S}\sum_{j=1}^{N} \|C_L(E_D(x_{i,j})) - \frac{1}{2})\|_2^2. \tag{4}$$

With facial content and domain losses, together with the corresponding confusion losses, our proposed framework allows us to disentangle liveness-irrelevant features from the input images. The deployment of E_C, E_D, D_C, and C_D would also facilitate the learning of liveness representation, as we discuss next.

Algorithm 1: Learning of our framework for domain-generalized facial liveness representation and FAS

```
Require:
  Input: face images $X$, binary labels $Y$, liveness encoder $E_L$, content encoder $E_C$, domain encoder $E_D$, liveness classifier $C_L$, content decoder $D_C$, domain classifier $C_D$
  while not done do
    Calculate objectives from Eq. (1) to Eq. (6)
    Disentangle Facial Content Representation
      Update $E_C$ with $L_{cont} + L_{cont}^{cnf}$
      Update $D_C$ with $L_{cont}$
    Disentangle Domain Representation
      Update $E_D$ with $L_{dom} + L_{dom}^{cnf}$
      Update $C_D$ with $L_{dom}$
    Learning of Generalized Liveness Representation
      Update $E_L$ with $L_{live} + L_{live}^{cnf}$
      Update $C_L$ with $L_{live}$
  end while
  return $E_L, C_L$
```

Learning Domain-Invariant Liveness Representation

To address domain-generalized FAS, learning of domain-invariant liveness representation from the input images would be the major component of our proposed framework. With facial content and image domain features properly disentangled from the input image, we now discuss how the extraction of liveness representation is realized by our network so that the derived features can be applied to detect novel spoof attacks in unseen target domains.

As depicted in Fig. 1, we have liveness encoder E_L and the associated classifier C_L deployed in our framework. Given the jth training image from source domain i, we have the encoded liveness feature $E_L(x_{i,j})$ expected to describe the liveness (i.e. real/fake) information. The subsequent liveness classifier C_L is designed to determine whether the input face image is real or fake depending on the feature $E_L(x_{i,j})$. To better separate real and fake facial images, we adopt the simplified Large Margin Cosine Loss (LMCL) function (Wang et al. 2018) as the objective, which calculates intra/inter-class angular distances for the corresponding input images with a predetermined margin m. For the sake of clarity, we disregard the domain index i for the input image; thus, *liveness loss* L_{live} for updating E_L and C_L is calculated as follows:

$$L_{live} = -\sum_{j=1}^{N} log(\frac{e^{\alpha(W_{y_j}^T E_L(x_j)-m)}}{e^{\alpha(W_{y_j}^T E_L(x_j)-m)} + e^{\alpha(W_{1-y_j}^T E_L(x_j))}}), \tag{5}$$

where α is a hyperparameter, and $W = \{W_0, W_1\}$ denotes the parameters of liveness classifier C_L. We note that, W_0 represents the model parameter for label $y_j = 0$ (i.e., spoof attack), while W_1 denotes that for $y_j = 1$ (i.e., real face). And, the margin m is introduced in Eq. (5), so that the separation between liveness representations derived from real and fake images can be further enforced.

To further ensure that the liveness feature $E_L(x_{i,j})$ does not contain any image domain information, we utilize the aforementioned domain classifier C_D, and calculate the *liveness confusion loss* L_{live}^{cnf} as follows:

$$L_{live}^{cnf} = \sum_{i=1}^{S}\sum_{j=1}^{N} \|C_D(E_L(x_{i,j})) - \frac{1}{S})\|_2^2. \tag{6}$$

With both liveness classification and confusion losses, we are able to train our proposed framework for disentangling domain-invariant liveness representation for domain-generalized FAS.

Together with Eq. (5) and Eq. (6), we maximize $W_{y_j}^T E_L(x_j)$ and minimize $W_{1-y_j}^T E_L(x_j)$, which encourages the separation between real and fake images (with margin m) across domains. On the other hand, maximization of $W_{y_j}^T E_L(x_j)$ further implies the suppression of intra-class variation for images of the same label. In other words, the learned W_1 and W_0 would represent the *prototypes* of domain-invariant liveness representations of real and fake images, respectively. Therefore, with the disentanglement of image domain information, our learning of liveness representation would not only separate real face images and spoof attacks but also enforce the minimization of the associated

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	50.3	51.64	29.76	78.5	54.28	44.98	50.29	49.31
Binaray CNN	34.47	65.88	29.25	82.87	34.88	71.94	29.61	77.54
IDA	28.35	78.25	66.67	27.86	55.17	39.05	54.20	44.59
Color Texture	40.40	62.78	28.09	78.47	30.58	76.89	63.59	32.71
LBPTOP	49.45	49.54	36.90	70.80	42.60	61.05	53.15	44.09
Auxiliary(Depth Only)	29.14	71.69	22.72	85.88	33.52	73.15	30.17	77.61
Auxiliary(All)	27.60	-	-	-	28.40	-	-	-
MMD-AAE	31.58	75.18	27.08	83.19	44.95	58.29	40.98	63.08
MADGG	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
RFM	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
SSDG-M	16.67	90.47	23.11	85.45	18.21	94.61	25.17	81.83
SSDG-R	**7.38**	97.17	10.44	95.94	11.71	**96.59**	**15.61**	91.54
Cross	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
DRDG	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.16
Ours	7.50	**97.45**	**9.80**	**96.82**	**11.38**	94.90	16.70	**91.83**

Table 1: Comparisons of FAS in unseen domains in terms of HTER and AUC. For example, O&C&I to M denotes that the model is trained on the datasets of Oulu, CASIA, & Idiap, and is evaluated on MSU. Note that the attack types are print and replay in both source and target domains.

intra-class variations. Therefore, the generalization of our model to novel spoof attacks across different domains can be expected.

Overall Objectives

With the introduced objectives discussed above, we now detail the training process of our proposed network architecture. Given training image samples from S source domains, we train our face content encoder E_C and decoder D_C with L_{cont} for deriving liveness-irrelevant content features, while L_{cont}^{cnf} is calculated from the outputs of liveness classifier C_L and domain classifier C_D for preserving the content-only information in the derived feature. As for the domain encoder E_D and classifier C_D, we calculate L_{dom} and additionally observe L_{dom}^{cnf} to ensure the learning of liveness-irrelevant domain features. Finally, the liveness encoder E_L and classifier C_L are trained by observing L_{live} and L_{live}^{cnf}, while the latter is utilized to ensure the disentangling of domain information from the resulting liveness representation. The aforementioned network modules are jointly trained (in an end-to-end learning fashion), and the pseudocode for training is summarized in Algorithm 1.

Once the training of our network architecture is complete, only E_L and C_L are utilized for performing domain-generalized FAS. That is, the liveness encoder E_L is applied to extract the domain-invariant liveness representation from the input image, which is fed into C_L for FAS prediction.

Experiment

Experimental Settings

Datasets. Five public face anti-spoofing datasets are utilized to evaluate the effectiveness of our method: OULU-NPU (Boulkenafet et al. 2017) (denoted as O), CASIA-FASD (Zhang et al. 2012) (denoted as C), Idiap Replay-Attack (Chingovska, Anjos, and Marcel 2012) (denoted as I), MSU-MFSD (Wen, Han, and Jain 2015) (denoted as M), and CelebA-Spoof (Zhang et al. 2020b) (denoted as Cb). The detailed statistics of each dataset are shown in Table 1. in the supplementary material.

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	58.80	54.52	42.17
Color Texture	55.17	47.89	53.31	45.16
LBPTOP	45.27	54.88	47.26	50.21
MADGG	41.02	64.33	39.35	65.10
SSDG-M	31.89	71.29	36.01	66.88
Cross	31.67	75.23	34.02	72.65
DRDG	31.28	71.50	33.35	69.14
Ours	**25.09**	**80.15**	**31.99**	**77.14**

Table 2: Comparisons of FAS in unseen domains using limited source domain data. Following (Shao et al. 2019; Jia et al. 2020; Wang et al. 2020; Liu et al. 2021), images from two source domains as selected for training, while a target domain unseen during training is applied for evaluation.

Implementation details. Our experiment was implemented in PyTorch. We normalize all the face images to $256\times256\times3$ pixels as the input of our proposed model. The output sizes of both the decoder D_C, and the PRNet are $64 \times 64 \times 1$ pixels. We set the batch size to 10 and utilized the AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate of 1.5e-4 for training. The proposed model is trained with 200 epochs. We have ResNet-18 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) for the encoders E_L, E_C and E_D, where the weights of the first layer are shared. The architectures of the liveness classifier C_L, content decoder D_C, and domain classifier C_D are detailed in Table 2. in the supplementary materials.

Method	O&C&I to Cb		O&M&I to Cb		O&C&M to Cb		I&C&M to Cb	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MADGG	42.46	62.26	44.96	57.01	50.08	57.18	48.92	51.86
RFM	41.49	60.19	43.32	65.96	42.83	59.37	35.93	67.32
SSDG-M	41.19	60.91	35.25	65.96	36.40	66.66	35.02	69.19
SSDG-R	20.29	86.87	**20.58**	86.54	25.05	82.11	19.86	88.58
Ours	**19.42**	**88.17**	20.60	**86.93**	**22.32**	**85.49**	**16.22**	**90.85**

Table 3: Comparisons of FAS in unseen domains with novel spoof attacks in terms of HTER and AUC. For example, O&C&I to Cb denotes that the model is trained on the datasets of Oulu, CASIA, & Idiap, and is evaluated on CelebA-Spoof. Note that the attack types are print and replay in the source domains, while Cb contains the spoof attack of 3D masks for testing.

Baseline and evaluation metrics. We compare our method to several state-of-the-art methods of FAS, including MS_LBP (Määttä, Hadid, and Pietikäinen 2011), Binary CNN (Yang, Lei, and Li 2014), Image Distortion Analysis (IDA) (Wen, Han, and Jain 2015), Color Texture (CT) (Boulkenafet, Komulainen, and Hadid 2016), LBP-TOP (de Freitas Pereira et al. 2014), Auxiliary (Wen, Han, and Jain 2015), MMD-AAE (Li et al. 2018), MADDG (Shao et al. 2019), RFM (Shao, Lan, and Yuen 2020), SSDG (Jia et al. 2020), Cross (Wang et al. 2020), and DRDG (Liu et al. 2021). Following (Shao et al. 2019; Shao, Lan, and Yuen 2020; Jia et al. 2020; Wang et al. 2020; Liu et al. 2021), we have the Half Total Error Rate (HTER) and Area under the Curve of ROC (AUC) for evaluation metrics in all experiments.

FAS in Unseen Target Domains

Followed (Shao et al. 2019; Shao, Lan, and Yuen 2020; Jia et al. 2020; Wang et al. 2020; Liu et al. 2021), we utilize four public datasets, i.e., O, C, M, and I, to evaluate the effectiveness of our model adapting to unseen datasets. We select three datasets as the source domains and the remaining one as the target domain.

As shown in Table 1, our approach achieved impressive results and performed against all the existing FAS approaches. This demonstrates that the liveness encoder E_L in our proposed model is able to extract generalized features for the unseen domains by disentangling the liveness-irrelevant information, including the domain features and the facial content. On the other hand, since previous works such as Binary CNN and Color Texture do not extract the domain information from the liveness features, they obtain inferior results on domain-generalized FAS where the target domains are not observed during training. While recent approaches including MADDG and SSDG consider extracting domain-generalized liveness features, the generalization ability of their methods is still limited. This is because that their extracted liveness features still contain facial information, which is irrelevant to liveness information.

To further verify the domain generalization capability of our method, we evaluate our model when only two source domains can be observed during training. Following (Shao et al. 2019; Jia et al. 2020; Wang et al. 2020; Liu et al. 2021), we have M&I as the source domains and select one dataset from C&O as the target domain. When trained under limited source domains, it is more challenging to perform domain generalized FAS since the model is more likely to over-fit to source datasets. As shown in Table 2, our approach outperforms all the existing FAS approaches by a wide margin in the two tasks. It is expected since our model puts more emphasis on the liveness-relevant feature of the face images with the assist of the facial content disentanglement. With the facial content such as the head poses various in each domain, the existing FAS approaches are not able to perform efficient training when the training data is insufficient. On the other hand, with the facial content disentangled, our model is able to extract generalized liveness features and thus achieves impressive results.

FAS with Novel Spoof Attacks in Unseen Target Domains

We evaluate the effectiveness of our model adapting to a real-world scenario where the model encounters unseen spoof attacks which are not observed during training. Five datasets are used in this setting: O, C, I, M, and Cb. We choose Cb as the target domain since Cb contains a unique attack type, i.e. 3D mask, which does not appear in other datasets. Three of the remaining datasets are selected as the source domains. Thus, we have four different tasks under this setting: O&C&I to Cb, O&M&I to Cb, O&C&M to Cb, and I&C&M to Cb.

As shown in Table 3, we can observe that our approach outperforms existing FAS approaches in all of the four tasks. Our proposed method can be better generalized to unseen spoof attacks because it learns the domain-invariant liveness representations, which are able to generalize to unseen target domains. To further enhance the generalization of our learned liveness representation, we disentangle the liveness-irrelevant facial content representation, from the liveness feature. Moreover, as explained in the *liveness loss* L_{live}, our liveness classifier C_L could learn a better liveness feature prototype after seeing different spoof attacks from the training data. Therefore, the learned fake prototype of the liveness classifier can generalize better to unseen spoof attacks. Previous DG approaches merely focused on extracting domain-invariant features, but they neglected the fact that their learned features still contained liveness-irrelevant information. When facing unseen spoof attacks, which are very different from their learned generalized feature for FAS, the generalization ability of their methods became very limited.

Method	**O&C&I to Cb**		**O&M&I to Cb**		**O&C&M to Cb**		**I&C&M to Cb**	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
Baseline	34.41	72.81	45.01	57.53	33.36	72.9	38.30	66.74
Baseline + E_C, D_C	33.26	73.23	41.94	61.63	28.13	77.35	39.15	68.78
Baseline + E_D, C_D	27.66	79.40	34.64	69.59	25.99	80.62	26.15	82.54
Ours	**19.42**	**88.17**	**20.60**	**86.93**	**22.32**	**85.49**	**16.22**	**90.85**

Table 4: Analysis of our network architecture design. Note that Baseline denotes the learning of only the liveness encoder and liveness classifier. The modules for disentangling liveness-irrelevant representations (i.e., E_C&D_C for content and E_D&C_D for image domain) are assessed, which are shown to support the effectiveness of our full model.

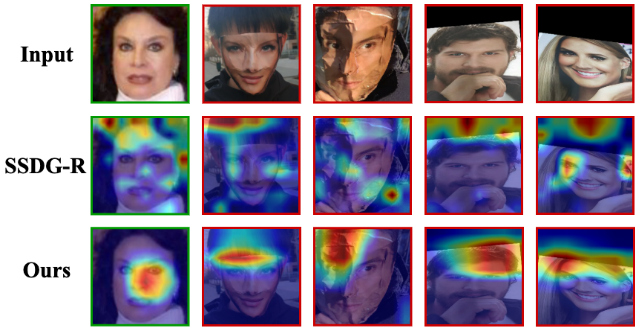

Figure 2: FAS visualization examples on real and spoof attacks (in green and red bounding boxes, respectively) using O&M&I to Cb. Note that the Grad-CAM is utilized to visualize the activation maps for real and fake images. Comparing to SSDG-R, our method offers proper attention and thus explains of why the input image contains a real or fake face.

Ablation Study

We provide the ablation study to verify the effectiveness of each component of our proposed method. As listed in Table 4, the baseline model (in the first row) contains only the liveness encoder E_L and liveness classifier C_L. The second and the third row shows the results when the facial content disentanglement and domain disentanglement is applied, respectively. The last row shows the results of our proposed model, i.e., both content and domain disentanglement are applied.

We can see from the first two rows that the model with the disentanglement of facial content performs better than the baseline. This confirms that such disentanglement helps our model to focus on the liveness cues in the images instead of liveness-irrelevant facial content. Comparing the first row to the third row, we see that the domain disentanglement mechanism helps our model extract domain-invariant liveness features that can be applied on the unseen target domain and therefore brings significant improvements. With both facial content and domain feature disentanglement are utilized, our proposed method achieved the best HTER and AUC in all tasks.

Qualitative Analysis

To verify the effectiveness of our representation disentanglement model, we utilize the Grad-CAM (Selvaraju et al. 2017) to obtain the class activation map visualizations, where the activation map indicates the regions that the model attends on when performing domain-generalized FAS.

We compare our model to SSDG-R and show the visualization results in Figure 2. The first column of the figure shows the visualization result of a real face and the other columns present the results of spoof faces. In the first column, SSDG-R focused on the background of the face image while our proposed model concentrated on the liveness part of the face image. For the spoof image in the third column, with the facial content features disentangled, our proposed model concentrated on the sunken part of the 3D mask, which could be regarded as important liveness information. On the other hand, SSDG-R focuses more on the liveness-irrelevant facial contours and background. The visualization results support the effectiveness of our representation disentanglement model for domain-generalized FAS.

Conclusion

In this paper, we address the challenging task of domain-generalized FAS problem, in which novel spoof attacks in unseen target domains need to be identified. Based on the idea of representation disentanglement, we present a network architecture that is able to extract facial liveness, content, and domain features. Aiming at performing domain-generalized FAS, the facial liveness representation exhibits domain invariant properties. On the other hand, the content and domain representations are viewed as liveness-irrelevant features, whose derivations are enforced by our network module and objective designs. Extensive experiments on benchmark datasets demonstrated the effectiveness of our proposed network, which shows promising domain generalization in addressing FAS, including the ability to handling novel types of spoof attacks during inference. Since the current proposed model only observes images as inputs, taking video inputs and thus dealing with visual-temporal information would be among our future research directions for domain-generalized FAS.

References

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

Boulkenafet, Z.; Komulainen, J.; and Hadid, A. 2016. Face spoofing detection using colour texture analysis. *IEEE*

Transactions on Information Forensics and Security, 11(8): 1818–1830.

Boulkenafet, Z.; Komulainen, J.; Li, L.; Feng, X.; and Hadid, A. 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)*, 612–618. IEEE.

Chingovska, I.; Anjos, A.; and Marcel, S. 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)*, 1–7. IEEE.

de Freitas Pereira, T.; Anjos, A.; De Martino, J. M.; and Marcel, S. 2012. LBP- TOP based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, 121–132. Springer.

de Freitas Pereira, T.; Anjos, A.; De Martino, J. M.; and Marcel, S. 2013. Can face anti-spoofing countermeasures work in a real world scenario? In *2013 international conference on biometrics (ICB)*, 1–8. IEEE.

de Freitas Pereira, T.; Komulainen, J.; Anjos, A.; De Martino, J. M.; Hadid, A.; Pietikäinen, M.; and Marcel, S. 2014. Face liveness detection using dynamic texture. *EURASIP Journal on Image and Video Processing*, 2014(1): 1–15.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Feng, L.; Po, L.-M.; Li, Y.; Xu, X.; Yuan, F.; Cheung, T. C.-H.; and Cheung, K.-W. 2016. 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.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 534–551.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jia, Y.; Zhang, J.; Shan, S.; and Chen, X. 2020. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8484–8493.

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

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5400–5409.

Liu, S.; Zhang, K.-Y.; Yao, T.; Sheng, K.; Ding, S.; Tai, Y.; Li, J.; Xie, Y.; and Ma, L. 2021. Dual reweighting domain generalization for face presentation attack detection. *arXiv preprint arXiv:2106.16128.*

Liu, Y.; Jourabloo, A.; and Liu, X. 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.

Liu, Y.; Stehouwer, J.; and Liu, X. 2020. On disentangling spoof trace for generic face anti-spoofing. In *European Conference on Computer Vision*, 406–422. Springer.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101.*

Määttä, J.; Hadid, A.; and Pietikäinen, M. 2011. Face spoofing detection from single images using micro-texture analysis. In *2011 international joint conference on Biometrics (IJCB)*, 1–7. IEEE.

Patel, K.; Han, H.; and Jain, A. K. 2016a. Cross-database face antispoofing with robust feature representation. In *Chinese Conference on Biometric Recognition*, 611–619. Springer.

Patel, K.; Han, H.; and Jain, A. K. 2016b. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10): 2268–2283.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shao, R.; Lan, X.; Li, J.; and Yuen, P. C. 2019. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10023–10031.

Shao, R.; Lan, X.; and Yuen, P. C. 2017. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3D mask face anti-spoofing. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 748–755. IEEE.

Shao, R.; Lan, X.; and Yuen, P. C. 2020. Regularized Fine-Grained Meta Face Anti-Spoofing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 11974–11981.

Wang, G.; Han, H.; Shan, S.; and Chen, X. 2020. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6678–6687.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wen, D.; Han, H.; and Jain, A. K. 2015. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4): 746–761.

Yang, J.; Lei, Z.; and Li, S. Z. 2014. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*.

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

Zhang, K.-Y.; Yao, T.; Zhang, J.; Tai, Y.; Ding, S.; Li, J.; Huang, F.; Song, H.; and Ma, L. 2020a. Face anti-spoofing via disentangled representation learning. In *European Conference on Computer Vision*, 641–657. Springer.

Zhang, Y.; Yin, Z.; Li, Y.; Yin, G.; Yan, J.; Shao, J.; and Liu, Z. 2020b. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European Conference on Computer Vision*, 70–85. Springer.

Zhang, Z.; Yan, J.; Liu, S.; Lei, Z.; Yi, D.; and Li, S. Z. 2012. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, 26–31. IEEE.