LEARNING FACIAL LIVENESS REPRESENTATION FOR DOMAIN GENERALIZED FACE ANTI-SPOOFING

Anonymous ICME submission

ABSTRACT

Face anti-spoofing (FAS) 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.

Index Terms— Face anti-spoofing, domain generalization, representation disentanglement, deep learning

1. 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 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 [1, 2], image distortion [3], temporal variation [4], or deep semantic features [5, 6]. Although promising results have been obtained by these methods, to perform FAS on unseen image domains which are not observed during training remains a challenging task. The performance of these FAS methods trained from a particular source domain would drop dramatically when different backgrounds, subjects, or shooting devices are encountered in a different target domain of interest.

To address the domain shift problem mentioned above, studies have exploited auxiliary information, such as face depth [7], for distinguishing live and spoof faces. However, these approaches still have their limitations since they highly depend on the accuracy of estimated auxiliary information. Therefore, 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 [8] proposed a multi-adversarial deep domain generalization (MADDG) framework to derive domain-invariant feature spaces for real and fake images with a dual-force triplet mining constraint. Extended from MADDG, Jia [9] proposed a single-side domain generalization (SSDG) learning framework that groups spoof types across domains together with a triplet-mining algorithm for the purpose of domain generalization. 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 [10].

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 [10], aiming to obtain more discriminative liveness features by disentangling domain-invariant representations from an image. While [10] 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 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, and confirm that our method performs favorably against state-of-the-art approaches in detecting novel spoof attacks in unseen image domains.

2. PROPOSED METHOD

2.1. 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.

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.

2.2. 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 [11], which is known for face alignment 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 corresponding face alignment. 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:

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

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}:

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

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

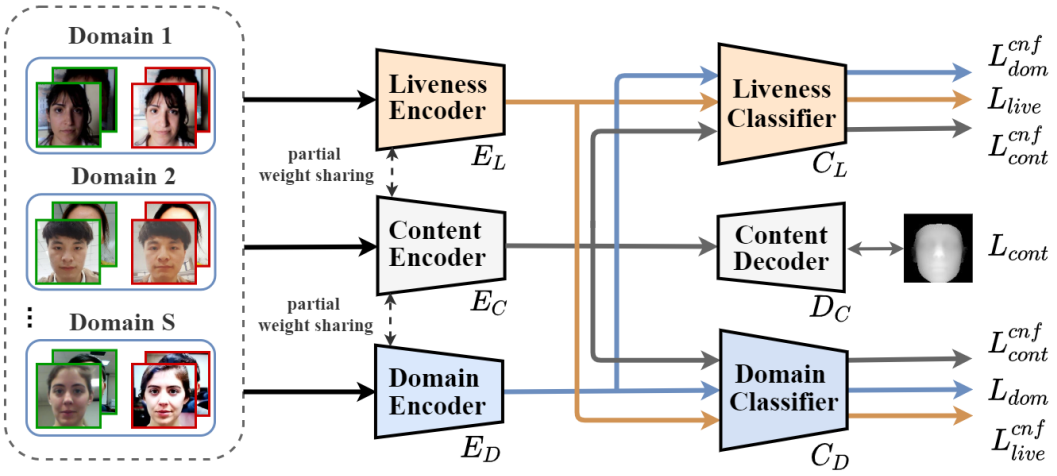

Fig. 1. Overview of our proposed network architecture. Our network aims to extract liveness, facial content, and image domain representations from facial input images. This is realized by the learning of liveness encoder E_L, content encoder E_C, and domain encoder E_D, respectively. To further ensure 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.

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:

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

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:

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

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.

2.3. 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 [12] 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:

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

where α is a hyperparameter, and $W = \{W_0, W_1\}$ denotes the parameters of liveness classifier C_L. We note that, W_0

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(%)
Auxiliary(Depth Only)	29.14	71.69	22.72	85.88	33.52	73.15	30.17	77.61
Auxiliary(All) [7]	27.60	-	-	-	28.40	-	-	-
MMD-AAE [13]	31.58	75.18	27.08	83.19	44.95	58.29	40.98	63.08
MADDG [8]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
RFM [14]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
SSDG-M [9]	16.67	90.47	23.11	85.45	18.21	94.61	25.17	81.83
SSDG-R [9]	**7.38**	97.17	10.44	95.94	11.71	**96.59**	**15.61**	91.54
Cross [10]	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
DRDG [15]	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.

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:

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

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

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

The overall objectives, together with the detailed training process, are summarized in the Algorithm 1 of our supplementary materials. 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, which is fed into C_L for FAS prediction.

3. EXPERIMENT

3.1. Experimental Settings

Five public face anti-spoofing datasets are utilized to evaluate the effectiveness of our method: OULU-NPU [16] (denoted as O), CASIA-FASD [17] (denoted as C), Idiap Replay-Attack [2] (denoted as I), MSU-MFSD [3] (denoted as M), and CelebA-Spoof [18] (denoted as Cb).

For the architecture of our model, we have ResNet-18 [19] pre-trained on ImageNet [20] for the encoders E_L, E_C and E_D, where the weights of the first layer are shared. The statistics of each dataset and the details of the architecture are shown in Table A. and Table B. of the supplementary materials, respectively. For the evaluation metrics, we have the Half Total Error Rate (HTER) and Area under the Curve of ROC (AUC) following [8, 9, 10, 15] in all experiments.

3.2. FAS in Unseen Target Domains

Following [8, 9, 10, 15], 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. 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.

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(%)
MADDG [8]	42.46	62.26	44.96	57.01	50.08	57.18	48.92	51.86
RFM [14]	41.49	60.19	43.32	65.96	42.83	59.37	35.93	67.32
SSDG-M [9]	41.19	60.91	35.25	65.96	36.40	66.66	35.02	69.19
SSDG-R [9]	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 2. Comparisons of FAS in unseen domains with novel spoof attacks in terms of HTER and AUC. Note that the attack types are print and replay in the source domains, while Cb contains the spoof attack of 3D masks for testing.

3.3. 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.

As shown in Table 2, 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. 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.

3.4. Qualitative Analysis

As discussed in the previous sections, recent DG approaches, including SSDG, focus on extracting domain-invariant features but neglect that the derived liveness features may still contain irrelevant facial content information. On the other hand, our method disentangles such facial content for better generalization performances. To verify the effectiveness of our disentanglement model, we utilize the Grad-CAM [21] algorithm 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 that disentanglement of the facial content enforce our model on the liveness-related part of the face images and therefore facilitate the domain-generalized FAS.

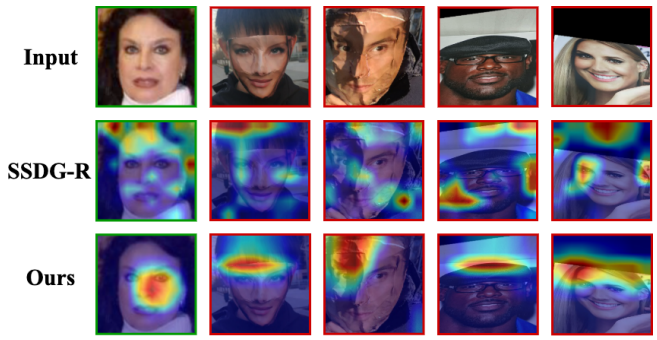

Fig. 2. FAS visualization examples on real and spoof attacks (in green and red bounding boxes, respectively) using O&M&I to Cb. Note that the Grad-CAM is utilized to visualize the activation maps for real and fake images. Comparing to SSDG-R, our method offers explainable attention maps.

3.5. Ablation Study

As listed in Table 3, 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 and third row, 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 performance in all tasks.

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 3. 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.

4. 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 while 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 handle 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.

5. REFERENCES

[1] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid, "Face anti-spoofing based on color texture analysis," in *2015 ICIP*. IEEE, 2015.

[2] Ivana Chingovska, André Anjos, and Sébastien Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG*. IEEE, 2012, pp. 1–7.

[3] Di Wen, Hu Han, and Anil K Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, 2015.

[4] 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 IJCB*. IEEE, 2017, pp. 748–755.

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

[6] Keyurkumar Patel, Hu Han, and Anil K Jain, "Cross-database face antispoofing with robust feature representation," in *Chinese Conference on Biometric Recognition*. Springer, 2016, pp. 611–619.

[7] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *CVPR*, 2018, pp. 389–398.

[8] Rui Shao et al., "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *CVPR*, 2019, pp. 10023–10031.

[9] Yunpei Jia et al., "Single-side domain generalization for face anti-spoofing," in *CVPR*, 2020, pp. 8484–8493.

[10] Guoqing Wang et al., "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *CVPR*, 2020, pp. 6678–6687.

[11] Yao Feng et al., "Joint 3d face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018, pp. 534–551.

[12] Hao Wang et al., "Cosface: Large margin cosine loss for deep face recognition," in *CVPR*, 2018, pp. 5265–5274.

[13] Haoliang Li et al., "Domain generalization with adversarial feature learning," in *CVPR*, 2018, pp. 5400–5409.

[14] Rui Shao, Xiangyuan Lan, and Pong C Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11974–11981.

[15] Shubao Liu et al., "Dual reweighting domain generalization for face presentation attack detection," *arXiv preprint arXiv:2106.16128*, 2021.

[16] Zinelabinde Boulkenafet et al., "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 612–618.

[17] Zhiwei Zhang et al., "A face antispoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*. IEEE, 2012, pp. 26–31.

[18] Yuanhan Zhang et al., "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *ECCV*. Springer, 2020, pp. 70–85.

[19] Kaiming He et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[20] Jia Deng et al., "Imagenet: A large-scale hierarchical image database," in *2009 CVPR*, 2009, pp. 248–255.

[21] Ramprasaath R Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.