Unsupervised Adversarial Domain Adaptation for Cross-Domain Face Presentation Attack Detection

Guoqing Wang, *Student Member, IEEE*, Hu Han, *Member, IEEE*, Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*— Face presentation attack detection (PAD) is essential for securing the widely used face recognition systems. Most of the existing PAD methods do not generalize well to unseen scenarios because labeled training data of the new domain is usually not available. In light of this, we propose an unsupervised domain adaptation with disentangled representation (DR-UDA) approach to improve the generalization capability of PAD into new scenarios. DR-UDA consists of three modules, i.e., ML-Net, UDA-Net and DR-Net. ML-Net aims to learn a discriminative feature representation using the labeled source domain face images via metric learning. UDA-Net performs unsupervised adversarial domain adaptation in order to optimize the source domain and target domain encoders jointly, and obtain a common feature space shared by both domains. As a result, the source domain PAD model can be effectively transferred to the unlabeled target domain for PAD. DR-Net further disentangles the features irrelevant to specific domains by reconstructing the source and target domain face images from the common feature space. Therefore, DR-UDA can learn a disentangled representation space which is generative for face images in both domains and discriminative for live vs. spoof classification. The proposed approach shows promising generalization capability in several public-domain face PAD databases.**

***Index Terms*— Face presentation attack detection, face liveness detection, face anti-spoofing, adversarial domain adaptation, metric learning, disentangled representation.**

Manuscript received October 30, 2019; revised April 13, 2020 and June 10, 2020; accepted June 10, 2020. Date of publication June 15, 2020; date of current version July 27, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700804 and in part by the Natural Science Foundation of China under Grant 61672496. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Corresponding author: Hu Han.)*

Guoqing Wang and Xilin Chen are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: guoqing.wang@vipl.ict.ac.cn; xlchen@ict.ac.cn).

Hu Han is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: hanhu@ict.ac.cn).

Shiguang Shan is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: sgshan@ict.ac.cn).

Digital Object Identifier 10.1109/TIFS.2020.3002390

I. Introduction

FACE recognition (FR) has been widely used in various applications such as smartphone unlock, access control, and pay-with-face. Since a genuine user's face image can be easily obtained by malicious users with a smartphone or from the social media, FR systems can be vulnerable to face presentation attacks (PA), e.g., printed photo, photo or video replay, and 3D mask [5]–[7]. Therefore, similar to biometric template protection [8], face PAD is a very important step for the existing FR systems, particularly for the face verification scenarios [9]–[11]. In recent years, a number of approaches have been proposed to detect various face presentation attacks, e.g., using hand-crafted features, deeply learned features, and auxiliary features.

Assuming that there are inherent disparities between live and spoof faces, most of the early PAD approaches utilized hand-crafted features to perform binary classification (live vs. spoof), e.g., using SVM classifiers [12]–[18]. These methods were found to be computationally efficient and work well under intra-database testing scenarios. With the success of deep learning such as Convolutional Neural Networks (CNNs) [19] in many computer vision tasks, recent PAD approaches seek to utilize CNNs for end-to-end face PAD or representation learning followed by binary classification models [20], [21]. For example, in [20], the features learned by deep learning show promising performance against the traditional hand-crafted feature based methods under intra-database testing scenarios. However, neither the hand-crafted feature based methods nor the deep feature based methods generalize very well to new application scenarios [11], [20] (see an example of deep feature learned for PAD in Fig. 1). The main reason is that the differences between live and spoof face images may include multiple aspects and different factors, such as skin detail, color distortion, moiré pattern, shape deformation, and texture artifacts. The presence of these factors in two domains (or databases) can be dramatically different; thus simply treating face PAD as a common two-class classification problem may not have good generalization ability. To improve the robustness of PAD methods under new scenarios, some scenario-invariant auxiliary information (such as face depth and heart rhythm) was leveraged to assist in live and spoof face classification [22], [23]. The performance of these methods

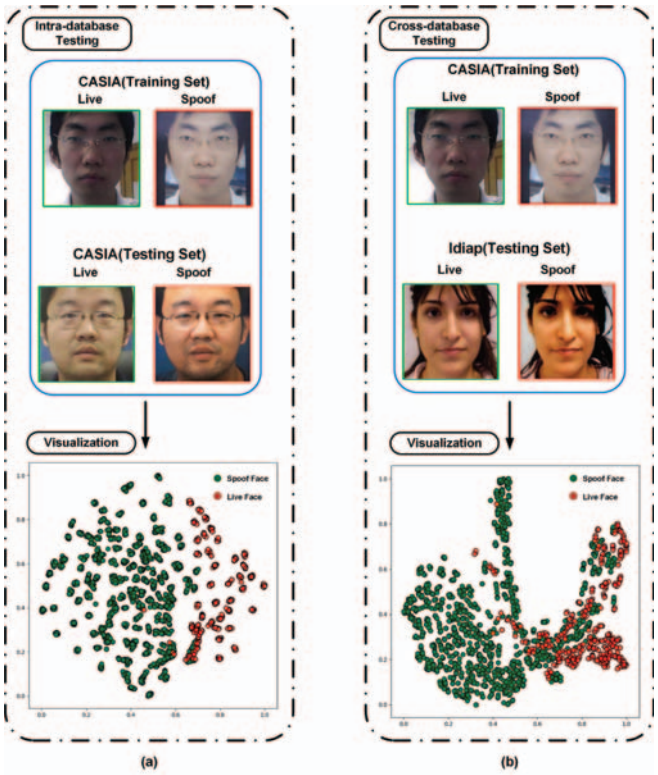

Fig. 1. The t-SNE [1] visualization of the features learned by ResNet-18 [2] for live and spoof face image classification on CASIA [3] and Idiap [4]. The model trained using the training set of CASIA is used for feature extraction on (a) the testing set of CASIA (intra-database testing), and (b) the testing set of Idiap (cross-database testing). We observe that such a straightforward model may not generalize well under cross-database testing scenario.

relies on the accuracy of the estimated auxiliary information to some extent.

Despite the tremendous progress on face PAD research, there are still limitations with existing methods: (i) While most face PAD approaches assume that the training and testing scenarios are similar in data distributions, there are often big disparities between them. This leads to poor generalization ability of PAD methods to practical application scenarios. (ii) There are various types of face presentation attacks; even for photo attack only, there can be printed photo, photo displayed on screen, etc. Therefore, it is not possible to build a labeled training set for each new application scenario, covering all possible presentation attacks. To address these issues, domain adaptation (DA) has been utilized to mitigate the gap between the target domain and the source domain during face PAD [24]–[26]. Recently, adversarial adaptation methods [25], [27] have sought to minimize an approximate domain discrepancy distance through an adversarial objective with respect to a domain discriminator because of the success of GAN [28].

In this paper, we focus on improving PAD generalization ability for cross-domain PAD[1] and propose a novel end-to-end trainable PAD approach called Unsupervised Domain Adaptation with Disentangled Representation (DR-UDA) to leverage the unlabeled data in the target domain and labeled data in the source domain to build a robust PAD model. For example, in FR based access control systems, a large number of face images can be recorded everyday, which may include both live face images and presentation attacks. However, these recorded face images are unlabeled. We aim to leverage such unlabeled face images in target domain and the labeled data in the source domains to build a PAD model that can generalize well to the target domain. To achieve this goal, we first learn a PAD model from the labeled source domain and then adapt it to the unlabeled target domain by learning a common feature space shared by both the source and target domains. The common feature space is expected to be discriminative for the live and spoof face images and generative for reconstructing the face images in both domains. The proposed approach is end-to-end trainable, and achieves promising results in cross-database face PAD on several public-domain databases (Idiap Replay-Attack (Idiap) [4], CASIA Face Anti-Spoofing (CASIA) [3], MSU-MFSD (MSU) [16], ROSE-YOUTU [24], CASIA-SURF [29] and Oulu-NPU (Oulu) [30], [31]).

The main contributions of this work are three-fold: (i) a novel unlabeled domain adaptation approach that is able to leverage labeled source domain data and unlabeled target domain data to build robust PAD model; (ii) disentangled representation learning for obtaining domain independent features during domain adaptation; and (iii) promising PAD performance in a number of cross-database tests than the state-of-the-art approaches addressing generalized face PAD.

Our preliminary work was described in [32]. Essential improvements in this work include: (i) we propose a disentangled representation learning, which allows more domain independent knowledge to be transferred and used for distinguishing live vs. spoof face images in the unlabeled target domain; (ii) we provide extensive evaluations using more datasets in public domain, e.g., Idiap, CASIA, MSU, ROSE-YOUTU, CASIA-SURF and Oulu databases, and provide comparisons with more baselines for PAD, such as DRCN [33], ADDA [25], DupGAN [34], Auxiliary [26], De-spoof [35] and STASN [36]; and (iii) we have provided more details about our method implementation, experimental evaluation, and related work.

II. Related Work

In this section, we provide a brief discussion of the related face PAD approaches, which can be generally categorized into hand-crafted feature based methods and deep feature based methods. We also briefly review the domain adaptation methods for PAD.

A. Face PAD

1) Hand-Crafted Feature Based Methods: Since most FR systems are using RGB sensors, print attack and replay attack become two major ways of presentation attacks. An early work by Li *et al.* [37] used Fourier spectra analysis to capture the difference between the live and spoof face images. After that, hand-crafted features such as LBP [15], LPQ [12], SURF [12], HoG [14], SIFT [18] and IDA [16] have been widely used with

[1]Cross-domain PAD means a PAD model is trained under one domain, and tested on a different domain.

traditional classifiers, such as SVM [42] and LDA [43], for a binary classification. To reduce the influence of illumination variations, some approaches converted face images from RGB color space to other space such as HSV and YCbCr [12], [38], and then extracted the above features for PAD. Besides extracting texture features, a number of methods also explored motion cues of the whole face or individual face components for PAD, such as eye blink [44], mouth movement [45] and head rotation [46]. Optical flow field was computed for discriminating between 2D planes (i.e., a printed photo) and 3D objects (i.e., a 3D live face).

The hand-crafted feature based PAD approaches are usually computationally efficient and explainable, and work well under intra-database testing scenarios. However, their generalization ability to new application scenarios is still not satisfying [20].

The background motion was also utilized for PAD in [47], [48]. Dynamic textures were considered in [7] to extract different facial motions. Liu *et al.* [5], [6] proposed to estimate rPPG signals from RGB face videos to detect attacks. Motion cue based PAD methods are expected to be more robust than the texture feature based methods under challenging illumination conditions; however, motion based methods often require users' cooperations, such as blinking eyes and rotating the head following specific instructions. Therefore, the system response time of these methods is usually longer than texture feature based PAD.

2) Deep Feature Based Methods: Heading into the era of deep learning, researchers are attempting to utilize deep models for face PAD [20], [21], [49] because of the great success of deep learning in many other computer vision tasks. In [21], ImageNet pretrained CaffeNet [50] and VGG-face [51] models were fine-tuned with live and spoof face images for face PAD. Xu *et al.* [52] adopted Long Short-Term Memory (LSTM) and CNN to obtain spatial-temporal features for PAD. Boulkenafet *et al.* [30] introduced a new challenging face PAD database namely Oulu-NPU and organized a face PAD competition [31]. Liu *et al.* [23] designed a novel framework to leverage the auxiliary information of depth and heart rate from the face videos to assist in PAD. Jourabloo *et al.* [35] inversely decomposed a spoof face into a live face and a noise of spoof, and then utilized the spoof noise for PAD. Wang *et al.* [53] recovered depth information from a temporal sequence, and used it for PAD. Chen *et al.* [54] proposed an attention-based method, which applies the complementary features (RGB and MSR) extracted via CNN and then employed the attention based fusion method to fuse these two features. Zhang *et al.* [29] introduced a large-scale multi-modal face anti-spoofing dataset namely CASIA-SURF, and proposed a multi-modality PAD framework which achieved higher performance than each single modality. Subsequently, in [55]–[57] several end-to-end approaches have been proposed to exploit the complementary information contained in RGB, depth and IR, and all reported the promising results on CASIA-SURF. Yang *et al.* [36] proposed a spatio-temporal attention mechanism to fuse global temporal and local spatial information for PAD. Wang *et al.* [58] proposed an effective disentangled representation learning for cross-domain presentation attack detection, which consists of a disentangled representation learning module and a multi-domain feature learning module.

While deep feature based PAD methods show strong feature learning ability, and can be trained end-to-end, there are still inherent constraints when the training set and testing set have a big discrepancy. As a result, the generalization ability of deep feature based PAD methods is still not satisfying under new application scenarios.

B. Domain Adaptation for PAD

Domain adaptation (DA) aims to transfer the knowledge or model learned from a source domain to a target domain [59]. DA can be very useful when there is only limited training data in a new application scenario, and thus has received increasing attention in recent years [60].

Long *et al.* [61] proposed a Deep Adaptation Network (DAN) to map deep features into Reproducing Kernel Hilbert Spaces (RKHS). Then, they performed DA by minimizing the maximum mean discrepancy (MMD) [62]. Muhammad *et al.* [33] proposed a deep reconstruction-classification network (DRCN) to learn a common representation for both domains through the joint objective of supervised classification of labeled source data and unsupervised reconstruction of unlabeled target data. A number of approaches have utilized adversarial learning proposed in Generative Adversarial Networks (GANs) [28] to reduce the source and target domain discrepancy for better DA [25], [34], [63], [64].

Face PAD also suffers from the cross-domain discrepancy issue, e.g., the distributions of data from the training and testing domains are different w.r.t. facial appearance, pose, illumination, sensor, etc. Therefore, recent studies for face PAD also attempted to utilize domain adaptation and domain generalization to overcome the poor generalization issues. Yang *et al.* [65] proposed a person-specific face anti-spoofing approach based on a subject-specific domain adaptation method to synthesize virtual features, which assumes that the relationship between live and spoof face images of the same subject can be formulated as a linear transformation. Li *et al.* [24] proposed an unsupervised domain adaptation PAD framework to transform the source domain feature space to the unlabeled target domain feature space by minimizing MMD. Shao *et al.* [41] proposed to learn a generalized feature space via a novel multi-adversarial discriminative deep domain generalization framework under a dual-force triplet-mining constraint.

A summary of the representative face presentation attack detection methods designed without domain adaptation and with domain adaptation are given in Table I. On the related note, there are some studies on PAD. Due to limited space, we refer interested readers to reviews in [9]–[11], [31]. While the prior work tried to utilize unlabeled target domain data to perform domain generalization, minimizing MMD alone may not fully exploit the useful information from the labeled source domain. In addition, the existing DA based face PAD methods do not consider how to enhance the domain independent feature learning given the labeled source domain data and the unlabeled target domain data.

TABLE I
SUMMARY OF THE REPRESENTATIVE FACE PRESENTATION ATTACK DETECTION METHODS WITH AND WITHOUT DOMAIN ADAPTATION

	Method	Feature	Auxiliary Info.	Loss Function
Methods w/o domain adaptation	Li et al. SPIE04 [37]	Fourier spectra from RGB image	-	-
	Chingovska et al. BIOSIG12 [4]	LBP from RGB image	-	-
	Wen et al. TIFS15 [16]	IDA feature from RGB image	-	-
	Boulkenafet et al. ICIP15 [38]	Color texture feature from RGB image	-	-
	Pinto et al. TIFS15 [39]	Time-spectral feature from RGB image	-	-
	Tirunagari et al. TIFS15 [40]	Dynamic feature from RGB image	-	-
	Yang et al. arXiv14 [21]	CNN feature from RGB image	-	• Cross-entropy Loss
	Liu et al. CVPR18 [23]	Depth and rPPG feature from RGB and HSV images	Depth + rPPG	• Depth Map Loss • rPPG Loss
	Jourabloo et al. ECCV18 [35]	Spoof noise feature from RGB and HSV images	Depth	• Magnitude Loss • Zero\One Map Loss • Repetitive Loss • GAN Loss
	Yang et al. CVPR19 [36]	Deep Spatio-Temporal feature from RGB image	-	• Cross-entropy Loss
Methods with domain adaptation	Li et al. TIFS18 [24]	Deep feature via unsupervised domain adaptation from RGB, HSV and YCbCr images	-	• MMD Loss
	Shao et al. CVPR19 [41]	Deep feature via adversarial multi-discriminative deep domain generalization from RGB and HSV images	Depth	• Depth Loss • Cls Loss • Triplet Loss • Adv Loss
	Wang et al. ICB19 [32]	Deep feature via adversarial domain adaptation from RGB image	-	• Triplet Loss • Adv Loss
	Proposed Method	Disentangled deep feature via adversarial domain adaptation from RGB image	-	• Triplet Loss • Center Loss • Adv Loss • Rec Loss

III. PROPOSED APPROACH

We aim to build a robust cross-domain face PAD model when there are only unlabeled face images in the new application domain and labeled live and spoof face images in the source domain. We propose a novel unsupervised adversarial domain adaptation method with disentangled representation (DR-UDA) to exploit the useful information from both domains. The overall framework of our DR-UDA is shown in Fig. 2. In DR-UDA, we aim to learn an effective face PAD model from the labeled source domain, and transfer it to the unlabeled target domain via a common feature space shared by both domains. The common feature space is expected to be discriminative for distinguishing between the live and spoof face images, but indiscriminative for distinguishing between the samples from two domains.

The proposed DR-UDA consists of a metric learning module (ML-Net), an unsupervised domain adaptation module (UDA-Net), and a disentangled representation learning module (DR-Net). We detail the three modules in the following sections.

A. ML-Net for Source Domain Metric Learning

Let X_s denote the source domain face images, and Y_s denote the corresponding labels (live or spoof). Let X_t denote the target domain face images without known labels. The ML-Net aims to learn a discriminative feature embedding ($\mathcal{F}$) using the labeled source domain face images (X_s, Y_s), i.e.,

$$f_{\theta_{S\_E}}(x) \in \mathcal{F}, \mathcal{F} \in \mathbb{R}^d. \tag{1}$$

where $x \in X_s$, $\mathcal{F}$ is a feature space that is expected to be discriminative for live and spoof face images. We use a deep residual neural network, i.e., ResNet-18 [2] with a new joint loss function to learn $f_{\theta_{S\_E}}(\cdot)$, the source encoder with parameter θ_{S_E}.

1) Joint Loss: Instead of using the common cross-entropy loss like previous face PAD methods [21], [52], we propose a novel joint loss consisting of triplet loss [66] and center loss [67] for metric learning to handle the big within-class diversity issue, particularly for the spoof face class, which consists of various spoof types. The center loss is defined as

$$\mathcal{L}_{center}(\theta_{S\_E}) = \frac{1}{2}\sum_{i=1}^{m} \left\| f_{\theta_{S\_E}}(x_i) - c_{y_i} \right\|_2^2, \tag{2}$$

where $c_{y_i} \in \mathbb{R}^d$ denotes the y_i-th class center in feature space. The triplet loss is defined as

$$\mathcal{L}_{triplet}(\theta_{S\_E}) = \sum_{(x_i^a, x_i^p, x_i^n)} max(\left\| f_{\theta_{S\_E}}(x_i^a) - f_{\theta_{S\_E}}(x_i^p) \right\|^2 - \left\| f_{\theta_{S\_E}}(x_i^a) - f_{\theta_{S\_E}}(x_i^n) \right\|^2 + \alpha, 0), \tag{3}$$

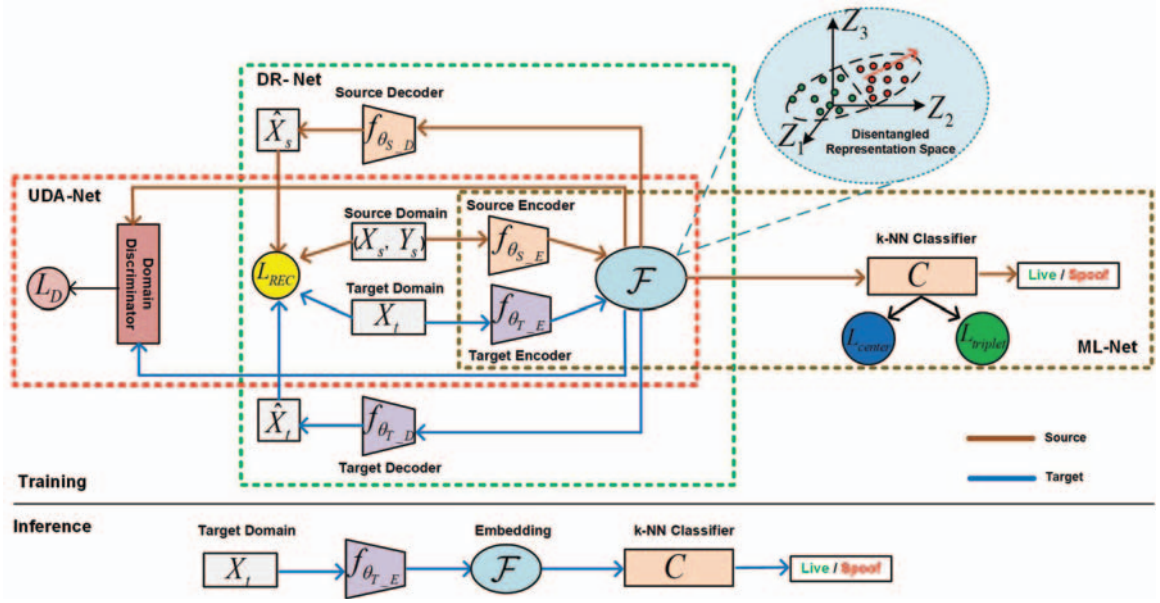

Fig. 2. The overall diagram of the DR-UDA framework. The DR-UDA consists of three modules, a source domain metric learning network (ML-Net), an unsupervised adversarial domain adaptation module (UDA-Net), and a disentangled representation learning module (DR-Net), respectively.

where x_i^a, x_i^p and x_i^n denote anchor sample, positive sample and negative sample, respectively. We want to ensure that an anchor data x_i^a, no matter belonging to live or spoof, can be closer to an image of the same class x_i^p than an image of different class x_i^n. Minimizing this term results in moving the anchor sample x_i^a towards positive samples x_i^p while pushing away from negative sample x_i^n in the embedding space. In addition, x_i^a is expected to be closer to x_i^p than to x_i^n by at least a margin of α in the embedding space. The final joint loss can be represented as

$$\mathcal{L}_{joint} = \mathcal{L}_{triplet} + \lambda \mathcal{L}_{center}, \tag{4}$$

where λ is a hyperparameter balancing the two losses, and we empirically set $\lambda = 1$ in our experiments. The live and spoof face images are expected to form two clusters in the embedding space, making it easy for discriminating them using k-nearest neighbors (k-NN) classifier [68] ($\mathcal{C}$). Compared to SVM [42] and FC layers [19], k-NN classifier can generate a highly convoluted decision boundary as it is driven by the raw training data itself.

2) Triplet Selection: In order to ensure good network convergence and effective representation learning, it is important to choose the appropriate triplets, particularly the hard ones. Given an anchor sample x_i^a, we determine its hard positive sample x_i^p by using the following rule

$$\arg\max_{x_i^p} \left\| f_{\theta_{S\_E}}(x_i^a) - f_{\theta_{S\_E}}(x_i^p) \right\|^2. \tag{5}$$

Similarly, we determine the hard negative sample x_i^n by using the following rule

$$\arg\min_{x_i^n} \left\| f_{\theta_{S\_E}}(x_i^a) - f_{\theta_{S\_E}}(x_i^n) \right\|^2. \tag{6}$$

Then, a triplet pair is composed of (x_i^a, x_i^p, x_i^n). Each triplet pair is used to compute the loss and update the parameters.

B. UDA-Net for Unsupervised Adversarial Domain Adaptation

Since the face images X_T in the target domain are assumed to be unlabeled, supervised learning in terms of live vs. spoof labels in the target domain is not possible. Therefore, we build a target encoder $f_{\theta_{T\_E}}(\cdot)$ that can leverage the source domain knowledge and the k-NN classifier learned from the source domain to effectively distinguish between live and spoof face images in the unlabeled target domain. We propose an unsupervised adversarial domain adaptation (UDA-Net) to learn this target domain encoder $f_{\theta_{T\_E}}(\cdot)$. We divide both the source domain feature learning model (f_{θ_S}) and target domain feature learning model (f_{θ_T}) into two parts: encoder and decoder (see Fig. 2). That is

$$\begin{aligned} f_{\theta_S}(x_s) &= \mathbb{E}_{x_s \sim X_s}[f_{\theta_{cls}}(f_{\theta_{S\_E}}(x_s))], \\ f_{\theta_T}(x_t) &= \mathbb{E}_{x_t \sim X_t}[f_{\theta_{cls}}(f_{\theta_{T\_E}}(x_t))], \end{aligned} \tag{7}$$

where the source encoder $f_{\theta_{S\_E}}$ and target encoder $f_{\theta_{T\_E}}$ are used to extract multi-scale features from face images. The k-NN classifier $f_{\theta_{cls}}$ utilizes the multi-scale features extracted by source encoder $f_{\theta_{S\_E}}$ to predict whether the input face image is live or spoof. Our UDA-Net aims to obtain a feature space which is shared by both the source and target domain encoders. Since this shared feature space is already discriminative for the live and spoof face images in the source domain because of the learning by ML-Net, we can expect it is also discriminative for the unlabeled face images in the target domain after domain adaptation. UDA-Net simultaneously optimizes the learning of $f_{\theta_{S\_E}}$ and $f_{\theta_{T\_E}}$ in an adversarial manner, i.e., the samples from the source domain are indistinguishable from the samples in the target domain in the shared feature space. To achieve this goal, we introduce a discriminator with adversarial loss in our UDA-Net (see Fig. 2).

1) Adversarial Loss: While $f_{\theta_{S\_E}}$ and $f_{\theta_{T\_E}}$ aim to obtain a shared feature representation space, in which the samples from the source and target domains are indistinguishable with each other, the discriminator D aims to separate the data from the two domains in the embedding space. Here, D is optimized using the conventional adversarial loss [28], i.e.,

$$\mathcal{L}_D(\theta_{T\_E}, \theta_D) = -\mathbb{E}_{x_s \sim X_s}[\log D_{\theta_D}(f_{\theta_{S\_E}}(x_s))] - \mathbb{E}_{x_t \sim X_t}[\log(1 - D_{\theta_D}(f_{\theta_{T\_E}}(x_t)))], \quad (8)$$

where θ_{S_E} and θ_{T_E} denote the parameters for the source and target encoders, respectively. θ_D denotes the parameters of the discriminator D_{θ_D}. We fix the parameters of source encoder learned by ML-Net when training UDA-Net and only optimize the parameters of $f_{\theta_{T\_E}}$ and D_{θ_D}.

We did not optimize UDA-Net by directly using the minimax loss. Instead, we split the objective into two independent objectives, i.e., one for optimizing encoders ($f_{\theta_{S\_E}}$ and $f_{\theta_{T\_E}}$), and the other for optimizing the discriminator (D_{θ_D}). The loss $\mathcal{L}_D$ for discriminator D_{θ_D} remains unchanged. The target encoder loss $\mathcal{L}_E$ is defined as

$$\mathcal{L}_E(\theta_{T\_E}, \theta_D) = -\mathbb{E}_{x_t \sim X_t}[\log D_{\theta_D}(f_{\theta_{T\_E}}(x_t))]. \quad (9)$$

In summary, the proposed unsupervised domain adaptation is optimized with the following two objective functions:

$$\min_{D_{\theta_D}} \mathcal{L}_D(\theta_{T\_E}, \theta_D) = -\mathbb{E}_{x_s \sim X_s}[\log D_{\theta_D}(f_{\theta_{S\_E}}(x_s))] - \mathbb{E}_{x_t \sim X_t}[\log(1 - D_{\theta_D}(f_{\theta_{T\_E}}(x_t)))], \quad (10)$$

and

$$\min_{f_{\theta_{T\_E}}} \mathcal{L}_E(\theta_{T\_E}, \theta_D) = -\mathbb{E}_{x_t \sim X_t}[\log D_{\theta_D}(f_{\theta_{T\_E}}(x_t))]. \quad (11)$$

C. DR-Net for Disentangled Representation Learning

Given ML-Net and UDA-Net, our core task is to learn a feature embedding ($\mathcal{F}$), which is shared by the source and target domains and discriminative for face images from both domains. However, since the target domain is unlabeled, unsupervised domain adaptation via UDA-Net alone may not fully exploit the domain-independent features that are useful for robust cross-domain PAD. In light of this, we further propose to disentangle the features that are irrelevant to specific domains (which can be considered as the noises) for cross-domain PAD from the shared feature space $\mathcal{F}$ by performing source and target domain reconstruction from the shared feature space $\mathcal{F}$. We denote such a disentangled representation learning module as DR-Net (see Figs 2 and 3).

Specifically, the reconstructions for the source and target domain from shared feature space $\mathcal{F}$ can be represented as

$$\hat{X}_s = \mathbb{E}_{x_s \sim X_s}[f_{\theta_{S\_D}}(f_{\theta_{S\_E}}(x_s))],\ f_{\theta_{S\_E}}(x) \in \mathcal{F},$$
$$\hat{X}_t = \mathbb{E}_{x_t \sim X_t}[f_{\theta_{T\_D}}(f_{\theta_{T\_E}}(x_t))],\ f_{\theta_{T\_E}}(x) \in \mathcal{F}, \quad (12)$$

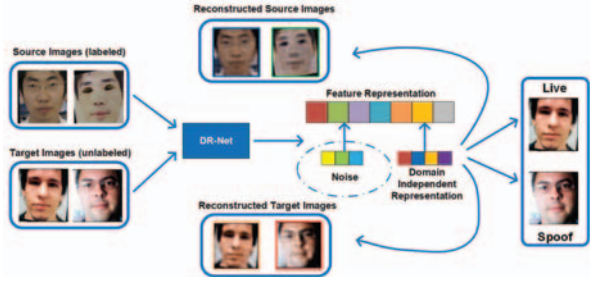

Fig. 3. With the face images from source and target domain as the inputs, DR-Net can disentangle the features irrelevant to specific domains (noises) from the common feature space learned by ML-Net and UDA-Net, and obtain a domain independent representation which is both generative and discriminative for cross-domain PAD.

where X_s and X_t denote the original face images in source and target domains, and $\hat{X}_s$ and $\hat{X}_t$ denote the corresponding reconstructed face images. $f_{\theta_{S\_D}}$ and $f_{\theta_{T\_D}}$ denote source and target decoders, respectively. To learn $f_{\theta_{S\_D}}$ and $f_{\theta_{T\_D}}$, and update the target encoder ($f_{\theta_{T\_E}}$) in DR-Net, we use a joint of L_1 loss and feature-level loss.

1) L_1 Loss: We use an L_1 loss to penalize the difference between the reconstructed face image and the input face image, i.e.,

$$\mathcal{L}_1(\theta_{S\_D}) = \mathbb{E}_{x_s \sim X_s}[\|x_s - f_{\theta_{S\_D}}(f_{\theta_{S\_E}}(x_s))\|_1],$$
$$\mathcal{L}_1(\theta_{T\_D}) = \mathbb{E}_{x_t \sim X_t}[\|x_t - f_{\theta_{T\_D}}(f_{\theta_{T\_E}}(x_t))\|_1]. \quad (13)$$

With L_1 loss, the two decoders $f_{\theta_{S\_D}}$ and $f_{\theta_{T\_D}}$ are expected to reconstruct the original source and target face images as accurate as possible.

2) Feature-Level Loss: Motivated by the success of the perceptual loss [69] used for image style transfer and super-resolution, we design a feature-level loss to enhance the training for DR-Net. Specifically, we use the source and target domain encoders $f_{\theta_{S\_E}}$ and $f_{\theta_{T\_E}}$ in ML-Net and UDA-Net to extract features from the original face images and reconstructed face images, which are then used for computing the MSE loss

$$\mathcal{L}_{fea}(\theta_{S\_D}) = \mathbb{E}_{x_s \sim X_s} \sum_{i=1}^{n} MSE[f^i_{\theta_{S\_E}}(x_s) - f^i_{\theta_{S\_E}}(f_{\theta_{S\_D}}(f_{\theta_{S\_E}}(x_s)))], \quad (14)$$

and

$$\mathcal{L}_{fea}(\theta_{T\_D}) = \mathbb{E}_{x_t \sim X_t} \sum_{i=1}^{n} MSE[f^i_{\theta_{T\_E}}(x_t) - f^i_{\theta_{T\_E}}(f_{\theta_{T\_D}}(f_{\theta_{T\_E}}(x_t)))], \quad (15)$$

where $f^i_{\theta_{S\_E}}(x)$ and $f^i_{\theta_{T\_E}}(x)$ represent the features from the convolutional layer before the i-th pooling layer extracted for image x by ResNet-18.

The overall objective function of DR-Net can be written as

$$\begin{aligned}\mathcal{L}_{REC}(\theta_{T\_E}, \theta_{S\_D}, \theta_{T\_D}) \\ = \mathcal{L}_{fea}(\theta_{S\_D}) + \lambda_1 \mathcal{L}_{fea}(\theta_{T\_D}) \\ + \lambda_2 \mathcal{L}_1(\theta_{S\_D}) + \lambda_3 \mathcal{L}_1(\theta_{T\_D}).\end{aligned} \tag{16}$$

We empirically set $\lambda_1 = 1$, $\lambda_2 = 2$ and $\lambda_3 = 10$ to balance the individual losses in our experiments.

We summarize the effectiveness of ML-Net, UDA-Net and DR-Net in our DR-UDA as follows. ML-Net aims to make the shared feature embedding $\mathcal{F}$ to be discriminative for the live vs. spoof face images in the source domain. UDA-Net aims to reduce the difference between the data distributions from source and target domain. With DR-Net included in DR-UDA, we can learn domain-independent features from $\mathcal{F}$ previously learned by ML-Net and UDA-Net for cross-domain PAD.

The proposed approach differs from [24] in that the way of domain adaptation between the source and target domains. [24] utilized the Maximum Mean Discrepancy between the source and target domains to perform domain adaptation. However, DR-UDA allows more domain independent knowledge to be transferred by adversarial and disentangled learning when using domain adaptation.

IV. Experimental Results

A. Databases

We provide evaluations on five public domain face databases for PAD including Idiap REPLAY-ATTACK [4], CASIA Face AntiSpoofing [3], MSU-MFSD [16], ROSE-Youtu [24] and Oulu-NPU [30]. We also use the RGB modality of the CASIA-SURF [29] dataset for cross-database evaluations.

1) Idiap REPLAY-ATTACK: Idiap REPLAY-ATTACK [4] (Idiap) consists of 1,200 videos from 50 subjects, which were taken by the webcam on a MacBook with the resolution of 320 × 240. The videos were captured under two conditions: (i) the controlled condition with uniform background and lighting, and (ii) the adverse condition with complex background and natural lighting. A Canon PowerShot camera was used to capture high-resolution face videos, which were then replayed using iPad 1 (1024 × 768) and iPhone 3GS (480 × 320), and printed on paper.

2) CASIA Face AntiSpoofing: CASIA Face AntiSpoofing Database [3] (CASIA) consists of 600 videos from 50 subjects, captured using multiple acquisition devices with different resolutions (i.e., Sony NEX-5 with the resolution of 1280× 720, and two webcams with the resolution of 640 × 480). The spoofing attacks include photo warping attack, cutting attack, and replay attack.

3) MSU-MFSD: MSU-MFSD [16] (MSU) consists of 280 videos from 35 subjects, which were captured using a Laptop camera and a smartphone camera with resolutions of 640 × 480 and 720 × 480, respectively. There are mainly two different spoofing attacks, e.g., printed photo attack and video replay attack.

TABLE II
Dataset Splits in Intra-Dataset PAD Tests

Dataset	train(#img.,#subj.)	test(#img.,#subj.)
Idiap [4]	(21600, 30)	(14352, 20)
CASIA [3]	(44705, 20)	(63882, 30)
MSU [16]	(4902, 15)	(5402, 20)
ROSE-Youtu [24]	(21320, 10)	(21417, 10)

4) ROSE-Youtu: ROSE-Youtu [24] consists of 4,225 videos of 20 subjects, which were captured using multiple acquisition devices with different resolutions (Hasee smartphone with resolution of 640 × 480, Huawei Smartphone with resolution of 640 × 480, iPad 4 with resolution of 640 × 480, iPhone 5s with resolution of 1280 × 720), and ZTE smartphone with resolution of 1280 × 720). There are mainly three different spoofing attacks, i.e., printed photo attack, video replay attack, and masking attack.

5) CASIA-SURF: CASIA-SURF [29] is a recently released multi-modal (RGB, Depth and IR) face database for PAD, which contains 1,000 Chinese subjects with each one live video and six fake videos per subject. The RGB, depth and infrared (IR) modalities were simultaneously captured using the Intel RealSense SR300 camera. The background area of the face was cropped from original videos to make the face PAD task more challenging. They choose one frame out of every ten frames after face detection, and split the database into training, validation and testing sets, which contain 300, 100, and 600 subjects and 148K, 48K, and 295K frames, respectively.

6) Oulu-NPU: Oulu-NPU [30] consists of 4950 real access and attack videos, which were recorded using the front cameras of six mobile devices (Samsung Galaxy S6 edge, HTC Desire EYE, MEIZU X5, ASUS Zenfone Selfie, Sony XPERIA C5 Ultra Dual and OPPO N3) in three sessions with different illumination conditions and background scenes (Session 1, Session 2 and Session 3). The presentation attack types considered in the OULU-NPU database are print and video replay attacks.

B. Implementation Details

1) Network Structure: The source and target encoders share the same structure. There are four residual blocks in the encoders and each block has four convolution layers, which have the same settings as the convolution part of ResNet-18 [2]. The input images in the source and target domain are mapped into 512-D embedding feature vectors in disentangled representation space by the source and target encoders. The source and target decoders have the same structure, i.e., with four transposed convolution layers to generate the reconstructed images of 248 × 248 × 3 from the 512-D embedding feature space. We use a kernel size of 4 for all transposed convolution layers and use ReLU [70] layer for activation. The domain discriminator consists of two output nodes and three fully connected layers with 256, 256, and

128 hidden units, respectively. Each of the first two layers uses ReLU for activation. Our method is implemented in PyTorch.

2) Training Details: We use an open source SeetaFace[2] algorithm to do face detection and landmark localization. All the detected faces are then normalized to 256×256 based on five facial keypoints (two eye centers, nose, and two mouth corners). We resize the cropped face region to the size of 248×248 and use an open source imgaug[3] library to perform data augmentation, i.e., random flipping, rotation, resizing, cropping and color distortion. The DR-UDA is trained end-to-end by minimizing the losses in Eqs. 4, 10, 11 and 16. To make the training manageable, we train our DR-UDA in three stages, by gradually increasing the employed constraints and target domain samples. We train 20, 50, and 50 epochs for the three stages, respectively. At stage-1, we use the source domain samples to train ML-Net optimized with the joint of center loss and triplet loss. We choose SGD with an initial learning rate of $1e^{-3}$, and a batch size of 64 to train the source encoder with fixed parameters used in the following training procedure. At stage-2, we additionally include the images from unlabeled target domain to continue the training of UDA-Net and choose Adam as the optimizers of domain discriminator and target encoder in UDA-Net with initial learning rates of $1e^{-3}$ and $1e^{-5}$, respectively. At stage-3, we update the parameters of DR-Net optimized with Adam with an initial learning rate of $1e^{-3}$ to obtain a more generative and discriminative feature representation.

C. Experimental Settings

We follow the state-of-the-art face PAD methods [24], [26], [32], [41], and report Half Total Error Rate (HTER) [71] in the cross-database testing. We perform cross-database testing on CASIA, Idiap, MSU and Rose-Youtu. We follow the same testing protocols as that used in [24], [32], i.e., train the model using all the images from dataset A, and test the model on a different dataset B (denoted as A→B). So, we have twelve cross-database tests in total: C→I, C→M, C→Y, I→C, I→M, I→Y, M→C, M→I, M→Y, Y→I, Y→C, Y→M, in which C, M, I and Y denote CASIA, MSU, Idiap and Rose-Youtu, respectively. We choose ML-Net which is the core module of DR-UDA as one of the baselines. A number of domain adaptation methods, such as DRCN [33], ADDA [25] and DupGAN [34] were engaged in eliminating the gap between source and target domain. Although these methods were not designed specifically for cross-database face PAD, they reported promising generalization ability in digital datasets. So we also use all these methods as our baselines and report their performance on face PAD. In addition, two state-of-the-art methods, i.e., KSA§ [24] and ADA [32] leveraged domain adaptation to improve the cross-database PAD performance, and reported promising results. So we also use them as the state-of-the-art baseline algorithms. There are also many approaches addressing generalized face PAD without using domain adaptation, such as Auxiliary [26], De-spoof [35] and STASN [36]. While these works have not reported all the results under the twelve tests, the comparisons could be included where available, e.g. commonly used C → I and I → C. To further demonstrate the effectiveness of our method, we also use the RGB images in CASIA-SURF [29] as a target domain dataset to evaluate the models learned on Idiap, CASIA, MSU, and Rose-Youtu.

In recent years, Oulu-NPU has been increasingly used in assessing the generalization of face PAD methods. We conduct two experiments to validate the effectiveness of the proposed method on Oulu. In the first experiment, we follow the same testing protocol as that used in [24], [32] and report the cross-database performance under O→I, O→M, O→C, I→O, M→O and C→O. Since MADDG [41] does not show how to train the model using only one source domain dataset, we only compare with [41] using the multiple source domain setting in the second experiment.

In intra-dataset PAD tests, where there is a standard testing protocol provided with the dataset (such as CASIA and Rose-Youtu database), we simply follow their protocols. For the other datasets (such as Idiap and MSU), we split the dataset according to the subjects. The dataset splits for individual datasets are given in Table II.

D. Cross-Database Testing

As shown in Table III, the methods like ADDA [25], DRCN [33], DupGAN [34] and KSA§ [24], which used domain adaptation do not always achieve higher performance compared to the results of ML-Net, Auxiliary [23], De-spoof [35] and STASN [36], which designed for generalized face PAD. For example, under the tests of C→I and I→C, the above DA methods perform worse than the Auxiliary [23], De-spoof [35] and STASN [36]. The possible reason is that these traditional DA methods are classified by simple FC layers optimized with cross-entropy loss. The generalization ability of the representation learned by source domain encoder is poor on the target domain, and thus the results after domain adaptation are still worse than those generalized face PAD methods. More favorably, ML-Net is optimized using the joint of center loss and triplet loss instead of cross-entropy loss, which may benefit from metric learning. We can observe that Auxiliary [23] and STASN [36] achieve the relatively stable and good results under C→I and I→C. This indicates that adding temporal information can make the model more robust for cross-database testing. The baseline method KSA§ [24] works better than ML-Net for most of the twelve tests. This indicates the usefulness and necessity of using DA for cross-database PAD. Our preliminary work ADA [32] performs better than KSA§ [24] under seven of the twelve tests. This suggests that adversarial learning is more effective than Maximum Mean Discrepancy (MMD) for DA, while both are unsupervised DA methods. As shown in Table IV, the proposed method achieves better average testing performance than both KSA§ [24] and our preliminary method ADA under three of four datasets which shows that our domain adaptation approach is able to learn representation with better generalization ability. This suggests that the proposed approach has big potential for PAD under new application scenarios. However, the proposed

[2]https://github.com/seetaface/SeetaFaceEngine

[3]https://github.com/aleju/imgaug

TABLE III
CROSS-DATABASE FACE PAD PERFORMANCE (HTER IN %) OF THE PROPOSED APPROACH AND THE BASELINE APPROACHES

Method	C → I	C → M	C → Y	I → C	I → M	I → Y	M → C	M → I	M → Y	Y → C	Y → I	Y → M	Avg
ADDA [25]	41.8	36.6	31.4	49.8	35.1	50.0	39.0	35.2	38.7	28.7	34.6	33.4	37.8
DRCN [33]	44.4	27.6	32.5	48.9	42.0	50.0	28.9	36.8	39.4	32.3	37.4	37.2	38.1
DupGAN [34]	42.4	33.4	30.8	46.5	36.2	47.0	27.1	35.4	34.5	24.6	35.9	33.4	35.6
Auxiliary [23]	27.6	-	-	28.4	-	-	-	-	-	-	-	-	-
De-spoof [35]	28.5	-	-	41.1	-	-	-	-	-	-	-	-	-
STASN [36]	31.5	-	-	30.9	-	-	-	-	-	-	-	-	-
KSA§ [24]	39.3	15.1	31.6	**12.3**	34.9	40.1	**9.1**	33.3	30.4	30.1	38.8	26.1	28.4
ADA [32]	17.5	9.3	29.4	41.5	30.5	41.7	17.7	5.1	32.7	34.1	30.3	31.5	26.8
ML-Net	43.3	14.0	32.4	45.4	35.3	42.8	37.8	11.5	34.6	25.7	30.7	32.6	32.2
Proposed method	**15.6**	**9.0**	**28.0**	34.2	**29.0**	**39.8**	16.8	**3.0**	**29.7**	**17.9**	**23.7**	**24.4**	**22.6**

TABLE IV
AVERAGE CROSS-DATABASE TESTING PERFORMANCE (HTER IN %) ON FOUR DATASETS BY KSA§ [24], ADA [32] AND PROPOSED METHOD

Method	Testing Set			
	MSU	Idiap	CASIA	Rose-Youtu
KSA§ [34]	25.4	37.1	**17.2**	34.0
ADA [32]	23.8	17.6	31.1	34.6
Proposed method	**20.8**	**14.1**	23.0	**32.5**

TABLE V
P-VALUES OF T-TEST BETWEEN THE PROPOSED APPROACH AND THE BASELINE APPROACHES. BECAUSE THREE METHODS (AUXILIARY [23], DE-SPOOF [35], AND STASN [36]) ONLY REPORTED RESULTS UNDER TWO TEST SETTINGS, T-TEST IS NOT PERFORMED FOR THESE THREE METHODS FOR FAIR COMPARISONS WITH THE OTHER METHODS

Method	ADDA	DRCN	DupGAN	KSA§	ML-Net
p-value	0.0003	0.0003	0.0017	0.1930	0.0379

method may not work well under some cases. For example, when DR-UDA is trained on Idiap, MSU or Rose-Youtu, the HTER error for cross-database testing on CASIA remains high, i.e., the average cross-database testing HTER is more than 20% higher. The possible reason is that the diversity of the spoof attacks in Idiap, MSU MFSD and Rose-Youtu are relatively limited compared with those presented in CASIA.

In order to demonstrate the effectiveness of the proposed method, we also carry out a statistical t-test. We compare the results by our method and the results by the baseline methods in Table III, and computed the corresponding p-values, as shown in Table V. We notice that the p-values are smaller than 0.05 except for KSA§, which proves that our method and KSA§, which both used domain adaptation, perform better than the other baseline methods, and our method performs slightly better than KSA§.

In addition to the above cross-database tests, we also report cross-dataset PAD performance on CASIA-SURF. Since the cross-database PAD performance on CASIA-SURF was not reported in [24], we only compare our approach with ADDA [25], DRCN [33], DupGAN [34] and ADA [34]. As shown in Table VI, our method again achieves lower PAD error than the other state-of-the-art methods which suggests that the proposed DR-UDA has better generalization ability into new

TABLE VI
CROSS-DATABASE TESTING PERFORMANCE (HTER IN %) ON CASIA-SURF(DENOTED AS CS) DATABASE BY THE PROPOSED METHOD AND THE STATE-OF-THE-ART DOMAIN ADAPTATION PAD METHODS

Method	I → CS	M → CS	C → CS	Y → CS
ADDA [25]	46.5	50.0	34.6	50.0
DRCN [33]	49.8	49.4	38.6	47.8
DupGAN [34]	49.0	49.6	29.6	43.2
ADA [32]	46.5	45.4	30.2	38.4
Proposed method	**44.3**	**43.2**	**26.4**	**36.8**

scenarios. We observe that the models trained on the Idiap and MSU achieve higher error than those trained on the CASIA and Rose-Youtu when testing on CASIA-SURF. The possible reason is that the Idiap and MSU do not contain print attacks in which persons hold flat or cut photos. However, these attack types appear in CASIA and Rose-Youtu. In addition, all subjects in CASIA and Rose-Youtu are Asians, while there are no Asians in Idiap and a small portion of Asians in MSU.

We also provide evaluations by using one of the four datasets for testing and the combination of the other three datasets for training. The performance of the proposed approach and the state-of-the-art approaches are shown in Table VII (a). We can see that leveraging bigger training set can usually improve the cross-database testing performance of all the methods. This indicates that collecting a large PAD dataset is an important and effective way for improving the generalization ability of a PAD model. Again, the proposed approach achieves lower error than ADDA [25], DRCN [33], DupGAN [34] and ADA [32].

We also report cross-database PAD performance on Oulu-NPU, which is a challenging database for generalized PAD methods. For the cross-domain PAD method MADDG [41], since it does not show how to train the model using only one source domain dataset, we only compare with [41] using the multiple source domain setting (see Table VII (b)). We can see that the proposed method performs better than MADDG under three of the four tests. The possible reason is that the proposed method could learn a common feature space shared by both the source and target domains. However, MADDG aims to use the source domain data alone (w/o knowing target domain data) to learn a discriminative feature space, and

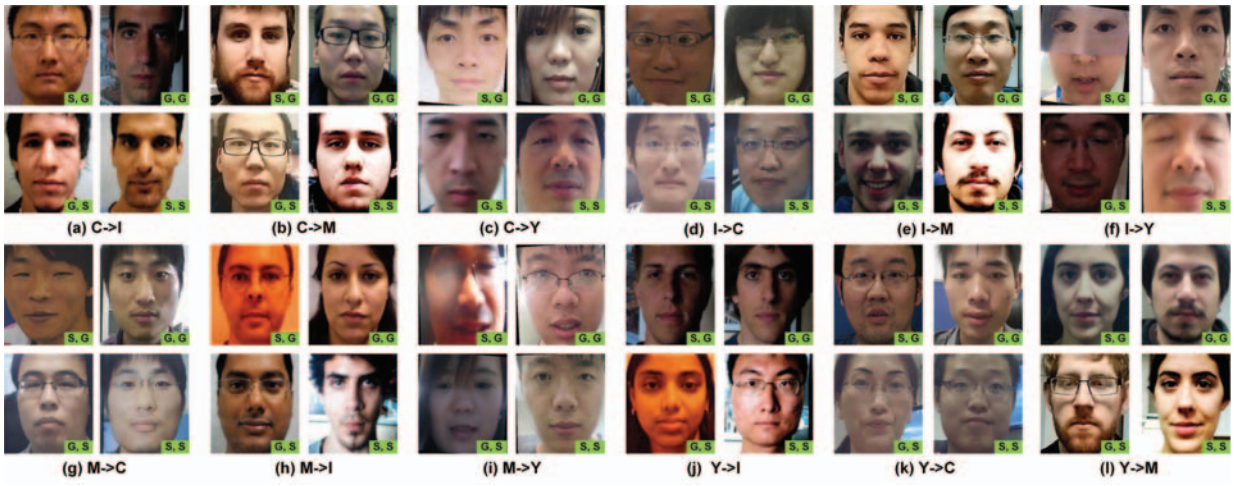

Fig. 4. Examples of correct and incorrect PAD results by the proposed approach in twelve cross-database tests(denoted as (a)-(i)). The label "S, G" (or "G, S") denotes a spoof (or genuine) face image is incorrectly classified as genuine (or spoof) face image; "G, G" (or "S, S") denotes a genuine (spoof) face image is correctly classified as genuine (spoof).

TABLE VII

CROSS-DATABASE TESTING PERFORMANCE (HTER IN %) OF INDIVIDUAL METHODS ON (A) IDIAP, MSU, CASIA, AND ROSE-YOUTU, AND (B) IDIAP, MSU, CASIA, AND OULU. IN EACH TEST, ONE OF THE FOUR DATASETS IS USED FOR TESTING, AND THE COMBINATION OF THE OTHER THREE DATASETS ARE USED FOR TRAINING

Method	[M, C, Y]→I	[I, C, Y]→M	[I, M, Y]→C	[I, C, M]→Y
ADDA [25]	35.4	34.2	33.4	36.3
DRCN [33]	37.2	33.9	31.4	35.7
DupGAN [34]	38.1	33.7	26.9	33.4
ADA [32]	6.3	12.7	37.8	31.0
Proposed method	**3.4**	**10.2**	**20.4**	**29.7**

(a)

Method	[O, C, I]→M	[O, M, I]→C	[O, C, M]→I	[I, C, M]→O
Aux(Depth) [23]	22.7	33.5	29.1	30.2
Aux(All) [23]	-	28.4	27.6	-
MADDG [41]	17.7	24.5	**22.2**	28.0
ADA [32]	16.9	24.2	23.1	25.6
Proposed method	**16.1**	**22.2**	22.7	**24.7**

(b)

there might be still domain gaps. We notice that the HTER of [M, C, Y]→I is much lower than [O, C, M]→I. The possible reason is that some live face images in Idipa have very similar background to the spoof face images in Oulu. We also notice that simply mixing multi-source datasets for training may lead to poor performance than using a single source dataset for training under some cases. The possible reasons are that: (1) the proposed approach may not benefit from a simply combined multi-source dataset unless each source dataset is specifically exploited for informative feature learning; and (2) severe data imbalance of the multi-source datasets may suppress the useful information contained in a small dataset. We also follow the same testing protocols as that used in [24], [32] and report cross-database PAD performance on Oulu-NPU (see Table VIII). We can observe that the proposed method performs better than the two baseline methods under all tests.

TABLE VIII

CROSS-DATABASE FACE PAD PERFORMANCE (HTER IN %) OF THE PROPOSED APPROACH AND THE BASELINE APPROACHES BETWEEN OULU AND OTHER THREE DATASETS (CASIA, IDIAP, AND MSU)

Method	O → I	O → M	O → C	I → O	M → O	C → O
CNN [25]	47.4	30.2	41.2	45.4	31.4	36.4
ADA [32]	26.8	31.5	19.8	29.6	31.2	29.1
ML-Net	40.1	32.8	38.7	44.2	31.6	34.3
Proposed	**25.4**	**27.4**	**19.5**	**38.5**	**30.2**	**28.7**

In this work, we assume that there are unlabeled face images containing both genuine faces and presentation attacks in the target domain. We agree that it is difficult to collect a new dataset containing both genuine and presentation attack samples (even unlabeled) for each new domain. However, we try to replicate the scenarios that when a face recognition system is used for access control, while most of the images are from genuine faces, there might be some chances (e.g., 25%) of either intentional presentation attacks or unintentional presentation attacks just for fun, collected by the face recognition system. We hope to adapt the pre-trained PAD model to the new scenario by using such unlabeled data collected by the face recognition system. This is the main motivation of unsupervised domain adaptation. Accordingly, in our experiments, we have verified the effectiveness of unsupervised domain adaptation when the percentages of live (spoof) samples vary among 0%, 25%, 50%, 75% and 100%. Fig. 6 shows the HTER changes of our approach when keeping the image number of face images of one class (either live or spoof) fixed, and changing the percentage of the other class. From Fig. 6 (a), we can notice that increasing the percentage of live face images does not bring too much gain in reducing the HTER. By contrast, we can notice from Fig. 6 (b) that increasing the percentage of presentation attack face images does bring more gain in reducing the HTER. The reason is that increasing the diversity of presentation attacks is important for improving cross-domain PAD robustness.

Fig. 4 shows some examples of PAD results by the proposed approach in the 12 cross-database tests. We notice that most

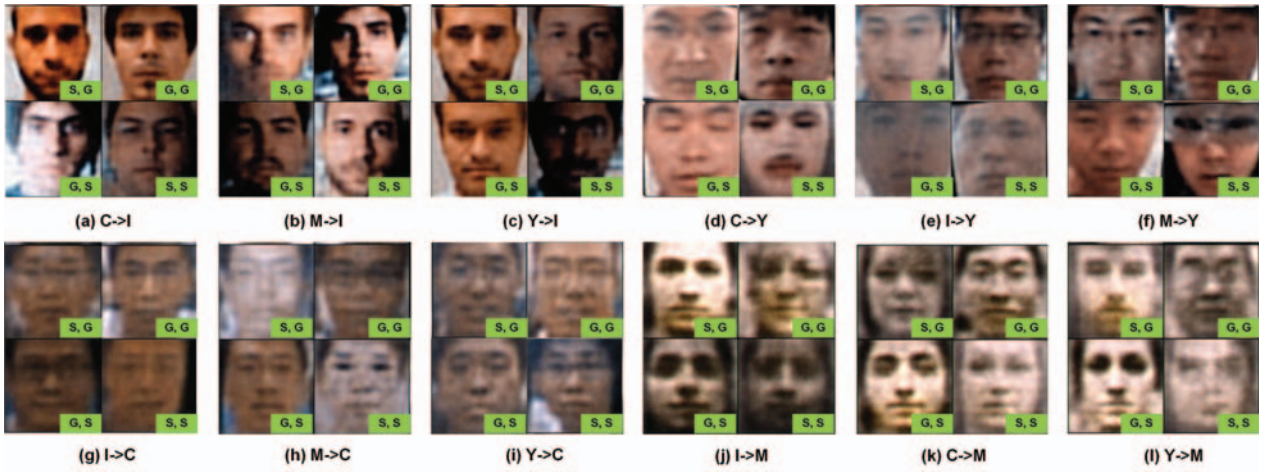

Fig. 5. Examples of reconstruction results of the target domain by the DR-Net in our DR-UDA in twelve cross-database tests from (a) to (i). The label "S, G" (or "G, S") denotes a spoof (or genuine) face image is incorrectly classified as genuine (or spoof) face image; "G, G" (or "S, S") denotes a genuine (spoof) face image is correctly classified as genuine (spoof).

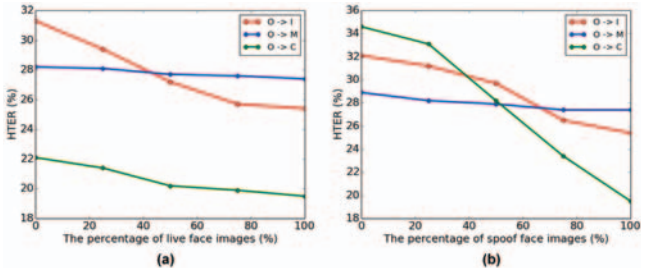

Fig. 6. The trend of HTER (in %) by DR-UDA when gradually enlarge the percentage of live (spoof) face samples while keeping the number of spoof (live) images fixed.

TABLE IX
INTRA-DATABASE TESTING PERFORMANCE (HTER IN %) OF THE PROPOSED METHOD AND THE BASELINE APPROACHES

Method	Idiap HTER	CASIA EER	MSU EER	Rose-Youtu EER
LPQ (HSV) [24]	7.9	7.4	12.2	30.4
LPQ (YCbCr) [24]	6.3	16.2	7.4	27.6
CoALBP (HSV) [24]	3.7	5.5	9.8	16.4
CoALBP (YCbCr) [24]	1.4	10.0	8.1	17.1
Deep learning [24]	2.1	7.6	5.8	8.0
ResNet-18 [2]	2.8	5.5	8.7	9.3
Proposed (ResNet-18)	1.4	**3.2**	**6.0**	**7.2**
SE-ResNet18 [72]	2.4	5.3	8.7	8.6
Proposed (SE-ResNet-18)	**1.3**	3.3	6.3	8.0

errors are caused by challenging appearance variances, such as over-saturated illumination, similar color distortions observed for both live and spoof face images, etc. In addition, the unseen attack types during training also lead to incorrect classification for the spoof face images in Figs. 4 (f) and (g).

Fig. 5 shows examples of reconstruction results by DR-Net in the target domain. We can observe that the reconstructed face images for Idiap look better than those on the other three databases. At the same time, our DR-UDA achieves lower PAD error on Idiap, as shown in Table IV. We observe that the learned common feature space is not overfitted to the source domain; instead, it is generative in retaining the semantic facial

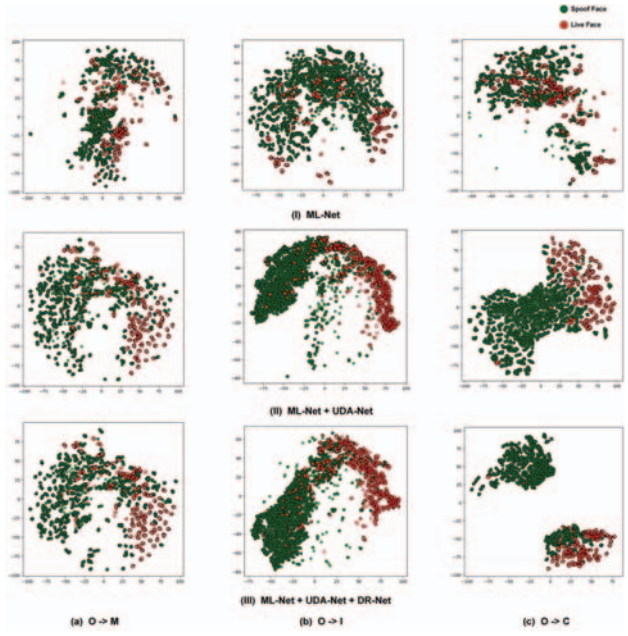

Fig. 7. The t-SNE [1] visualization of the live and spoof face images from MSU, Idiap, CASIA when source domain is Oulu in different feature spaces: (I) learned by ML-Net alone, (II) learned by the DR-UDA w/o DR-Net (ML-Net + UDA-Net), and (III) learned by the whole DR-UDA (ML-Net + UDA-Net + DR-Net).

information (shape, illumination, etc.) for the face images in the target domain. Such a capability is useful in improving the generalization ability of our PAD model into new domains.

E. Intra-Database Testing

Since many previous methods for face PAD only reported their performance under intra-database testing scenarios, we also perform intra-database testing on CASIA, Idiap and MSU, respectively. In [24], LPQ, CoALBAP, and deep learning features were reported to have the promising performance in intra-database testing. Therefore, we use the methods studied in [24] as our baselines for intra-database testing. Considering the success of deep learning in different tasks, we also

TABLE X

PERFORMANCE (HTER IN %) OF THE PROPOSED METHOD UNDER ABLATION STUDY IN TERMS OF METRIC LEARNING (ML), UNSUPERVISED DOMAIN ADAPTATION (UDA) AND DISENTANGLED REPRESENTATION LEARNING (DR)

Method	C → I	C → M	C → Y	I → C	I → M	I → Y	M → C	M → I	M → Y	Y → C	Y → I	Y → M
w/o ML&UDA& DR	43.8	33.8	40.3	49.5	41.3	50.0	45.7	39.6	45.7	35.2	37.6	45.6
w/o ML&UDA	44.4	33.6	34.5	48.9	42.0	50.0	43.2	40.2	41.7	34.3	37.2	40.0
w/o ML&DR	43.7	29.6	31.4	50.0	35.4	50.0	46.5	38.7	38.9	29.5	34.7	34.2
w/o UDA&DR	43.3	14.0	32.4	45.4	35.3	42.8	37.8	11.5	34.6	25.7	30.7	32.6
w/o ML	43.8	29.6	39.5	50.0	35.6	50.0	47.5	44.5	45.8	33.4	35.2	44.3
w/o UDA	43.3	14.3	34.5	45.3	34.8	50.0	38.2	12.4	40.4	33.2	36.3	39.4
w/o DR	17.5	9.3	29.4	41.6	30.5	41.7	17.7	5.1	34.1	18.7	30.7	31.5
Proposed method	**15.6**	**9.0**	**28.0**	**34.2**	**29.0**	**39.8**	**16.8**	**3.0**	**29.7**	**17.9**	**23.7**	**24.4**

TABLE XI

P-VALUES OF T-TEST FOR THE PROPOSED APPROACH UNDER ABLATION STUDY IN TERMS OF METRIC LEARNING (ML), UNSUPERVISED DOMAIN ADAPTATION (UDA) AND DISENTANGLED REPRESENTATION LEARNING (DR).

Method	w/o ML&UDA&DR	w/o ML&UDA	w/o ML&DR	w/o UDA&DR	w/o ML	w/o UDA	w/o DR
p-value	0.0000	0.0000	0.0003	0.0380	0.0000	0.0100	0.5119

use ResNet-18 [2] and SE-ResNet18 [72] to perform intra-database testing. For fair comparisons, we also use ResNet-18 and SE-ResNet-18 as the backbone network of the ML-Net in our DR-UDA during intra-database testing. We follow [24] and report HTER on Idiap, and EER on CASIA and MSU.

From the results in Table IX, we notice that the [24] using deep features, ResNet [2] and SE-ResNet [72] show stronger representation capacity than the hand-crafted features. We also notice that the ML module using ResNet-18 in our DR-UDA achieves much lower PAD errors than the baseline methods on CASIA, Rose-Youtu and MSU, while our ML module with 'SE-ResNet18' achieves the lowest PAD error on Idiap. These results suggest that the proposed approach is more effective in learning a discriminative feature representation for live vs. spoof face image classification. In addition, we can observe that the Squeeze-and-Excitation [72] is also effective in the PAD task.

F. Ablation Study

We provide ablation study to investigate the effectiveness of the three components in our DR-UDA, i.e., (i) metric learning for PAD via ML-Net, (ii) adversarial domain adaptation via UDA-Net, and (iii) disentangled representation learning via DR-Net. We study their influences by gradually dropping them from DR-UDA, and denote the corresponding models as 'w/o ML', 'w/o UDA', 'w/o DR', 'w/o ML&UDA', 'w/o ML&DR', 'w/o UDA&DR' and 'w/o ML&UDA&DR'.

The results of these models under cross-database testing on CASIA, Idiap, MSU and Rose-Youtu are given in Table X and Fig. 8. We can see dropping any of the three components will lead to increased PAD error. This suggests that all the three components are useful for our DR-UDA approach. In addition, we notice that ML-Net has a bigger influence than the other two modules to the generalization abilities of the proposed DR-UDA. For example, in the cross-database testing on Idiap, CASIA and MSU, removing the ML module from DR-UDA leads to 25%, 20% and 15% higher HTER than DR-UDA, respectively. The possible reason is that the UDA-Net and

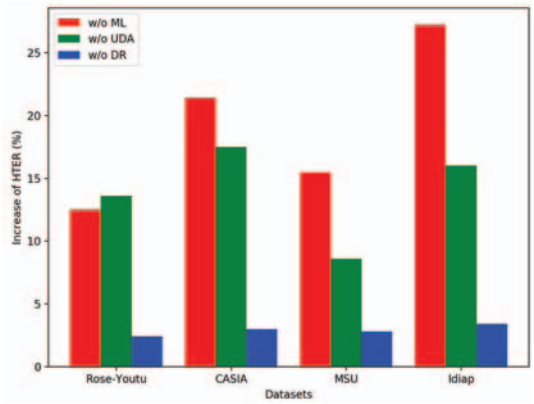

Fig. 8. The increase of average HTER (in %) by DR-UDA when dropping the ML-Net, UDA-Net or DR-Net module for ablation study.

DR-Net modules are fine-tuned in the feature space learned by ML-Net, and thus the feature space is primarily determined by ML-Net. Still, the UDA-Net and DR-Net can contribute to improving the cross-database testing performance. For example, if we visualize the feature representation learned by ML-Net alone, DR-UDA w/o ML-Net and the whole DR-UDA, we can see using UDA-Net and DR-Net together with ML-Net can obtain more robust feature representations that are discriminative for genuine vs. spoof face image classification (see Fig. 7). We also conduct statistical t-test for the proposed approach under ablation study. As shown in Table XI, we observe that the p-values of all tests are smaller than 0.05 except for 'w/o DR', which indicates that UDA-Net and ML-Net play more important roles than DR-Net. In addition, compared with the results by discarding all three components, the full method shows signification improvement in cross-dataset testing.

V. CONCLUSION

This paper addresses cross-domain face presentation attack detection (PAD)and proposes an unsupervised adversarial

domain adaptation (DR-UDA) method that can leverage unlabeled target domain data and labeled source domain data to build robust PAD model. DR-UDA consists of ML-Net, UDA-Net and DR-Net. ML-Net uses the combination of center loss and triplet loss jointly to learn a feature representation for live vs. spoof face image classification in the source domain. We then adapt this representation to the target domain via UDA-Net and DR-Net, so that this representation can be shared by both the source and target domains, and can be discriminative for live vs. spoof classification. The proposed approach outperforms the state-of-the-art face PAD methods on the a number of the public databases under the challenging cross-database testing scenario. Our future work includes utilizing 3D face prior knowledge and physiological cues to further improve the robustness of PAD models. In addition, we will study how to learn better representations that can further reduce the domain gap during domain adaptation.

REFERENCES

[1] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[3] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. ICB*, 2012, pp. 26–31.

[4] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. BIOSIG*, 2012, pp. 1–7.

[5] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3D mask face anti-spoofing with remote photoplethysmography," in *Proc. ECCV*, 2016, pp. 85–100.

[6] S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3D mask face presentation attack detection," in *Proc. ECCV*, 2018, pp. 558–573.

[7] R. Shao, X. Lan, and P. C. Yuen, "Joint discriminative learning of deep dynamic textures for 3D mask face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 923–938, Apr. 2019.

[8] K. Nandakumar and A. K. Jain, "Biometric template protection: Bridging the performance gap between theory and practice," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 88–100, Sep. 2015.

[9] J. Hernandez-Ortega, J. Fierrez, A. Morales, and J. Galbally, "Introduction to face presentation attack detection," in *Handbook of Biometric Anti-Spoofing*. New York, NY, USA: Springer, 2019, pp. 187–206.

[10] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[11] Z. Akhtar, C. Micheloni, and G. L. Foresti, "Biometric liveness detection: Challenges and research opportunities," *IEEE Secur. Privacy*, vol. 13, no. 5, pp. 63–72, Sep. 2015.

[12] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and Fisher vector encoding," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 141–145, Feb. 2017.

[13] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Lbp-top based countermeasure against face spoofing attacks," in *Proc. ACCV*, 2012, pp. 121–132.

[14] J. Komulainen, A. Hadid, and M. Pietikainen, "Context based face anti-spoofing," in *Proc. BTAS*, 2013, pp. 1–8.

[15] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Proc. IJCB*, 2011, pp. 1–7.

[16] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.

[17] K. Patel, H. Han, A. K. Jain, and G. Ott, "Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks," in *Proc. ICB*, 2015, pp. 98–105.

[18] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2268–2283, Oct. 2016.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[20] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Proc. CCBR*, 2016, pp. 611–619.

[21] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014, *arXiv:1408.5601*. [Online]. Available: http://arxiv.org/abs/1408.5601

[22] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. IJCB*, Oct. 2017, pp. 319–328.

[23] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. CVPR*, 2018, pp. 389–398.

[24] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 7, pp. 1794–1809, Jul. 2018.

[25] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, p. 4.

[26] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot, "Learning generalized deep feature representation for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 10, pp. 2639–2652, Oct. 2018.

[27] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. ICML*, 2018, pp. 1989–1998.

[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[29] S. Zhang *et al.*, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proc. CVPR*, 2019, pp. 919–928.

[30] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. FG*, 2017, pp. 612–618.

[31] Z. Boulkenafet *et al.*, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *Proc. IJCB*, 2017, pp. 688–696.

[32] G. Wang, H. Han, S. Shan, and X. Chen, "Improving cross-database face presentation attack detection via adversarial domain adaptation," in *Proc. ICB*, 2019, pp. 1–8.

[33] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. ECCV*. New York, NY, USA: Springer, 2016, pp. 597–613.

[34] L. Hu, M. Kan, S. Shan, and X. Chen, "Duplex generative adversarial network for unsupervised domain adaptation," in *Proc. CVPR*, 2018, pp. 1498–1507.

[35] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proc. ECCV*, 2018, pp. 290–306.

[36] X. Yang *et al.*, "Face anti-spoofing: Model matters, so does data," in *Proc. CVPR*, 2019, pp. 3507–3516.

[37] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of Fourier spectra," *Proc. SPIE*, vol. 5404, pp. 296–304, Aug. 2004.

[38] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *Proc. ICIP*, 2015, pp. 2636–2640.

[39] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4726–4740, Dec. 2015.

[40] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, "Detection of face spoofing using visual dynamics," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 762–777, Apr. 2015.

[41] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *Proc. CVPR*, 2019, pp. 10023–10031.

[42] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[43] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[44] L. Sun, G. Pan, Z. Wu, and S. Lao, "Blinking-based live face detection using conditional random fields," in *Proc. ICB*, 2007, pp. 252–260.

[45] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-time face detection and motion analysis with application in 'liveness' assessment," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 548–558, Sep. 2007.

[46] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *Proc. IASP*, 2009, pp. 233–236.

[47] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Proc. ICCV*, 2007, pp. 1–8.

[48] A. Anjos, M. M. Chakka, and S. Marcel, "Motion-based countermeasures to photo attacks in face recognition," *IET Biometrics*, vol. 3, no. 3, pp. 147–158, Sep. 2014.

[49] L. Feng *et al.*, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 451–460, Jul. 2016.
[50] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM MM*, 2014, pp. 675–678.
[51] O. M. Parkhi *et al.*, "Deep face recognition," in *Proc. BMVC*, 2015, vol. 1, no. 3, p. 6.
[52] Z. Xu, S. Li, and W. Deng, "Learning temporal features using LSTM-CNN architecture for face anti-spoofing," in *Proc. ACPR*, 2015, pp. 141–145.
[53] Z. Wang *et al.*, "Exploiting temporal and depth information for multi-frame face anti-spoofing," 2018, *arXiv:1811.05118*. [Online]. Available: http://arxiv.org/abs/1811.05118
[54] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 578–593, Jun. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8737949
[55] G. Wang, C. Lan, H. Han, S. Shan, and X. Chen, "Multi-modal face presentation attack detection via spatial and channel attentions," in *Proc. CVPRW*, 2019, p. 7.
[56] A. Parkin and O. Grinchuk, "Recognizing multi-modal face spoofing with face recognition networks," in *Proc. CVPRW*, 2019, p. 7.
[57] T. Shen, Y. Huang, and Z. Tong, "FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing," in *Proc. CVPRW*, 2019, p. 7.
[58] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proc. CVPR*, 2020, pp. 6678–6687.
[59] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2014, *arXiv:1409.7495*. [Online]. Available: http://arxiv.org/abs/1409.7495
[60] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," 2017, *arXiv:1702.05374*. [Online]. Available: http://arxiv.org/abs/1702.05374
[61] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015, *arXiv:1502.02791*. [Online]. Available: http://arxiv.org/abs/1502.02791
[62] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. NIPS*, 2016, pp. 136–144.
[63] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. NIPS*, 2016, pp. 469–477.
[64] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. NIPS*, 2017, pp. 700–708.
[65] J. Yang, Z. Lei, D. Yi, and S. Z. Li, "Person-specific face antispoofing with subject domain adaptation," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 797–809, Apr. 2015.
[66] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2004, pp. 234–778.
[67] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*, 2016, pp. 499–515.
[68] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, 2006, pp. 1473–1480.
[69] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*. New York, NY, USA: Springer, 2016, pp. 694–711.
[70] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.
[71] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: A public database and a baseline," in *Proc. IJCB*, 2011, pp. 1–7.
[72] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

Guoqing Wang (Student Member, IEEE) received the B.E. degree from North China Electric Power University in 2016. He is currently pursuing the joint M.S. degree with the University of Chinese Academy of Sciences, and the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics.

Hu Han (Member, IEEE) received the B.S. degree in computer science from Shandong University in 2005 and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), in 2011. He was a Research Associate with the PRIP Laboratory, Michigan State University, and a Visiting Researcher with Google, Mountain View. He joined the faculty at ICT, CAS, as an Associate Professor, in 2015. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics and medical image analysis. He has authored or coauthored over 60 articles in refereed journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, the IEEE TRANSACTIONS ON MEDICAL IMAGING, *Pattern Recognition*, CVPR, ECCV, NeurIPS, and MICCAI. He was a recipient of the IEEE FG2019 Best Poster Presentation Award and the CCBR 2016/2018 Best Student/Poster Awards. He is currently an Associate Editor of *Pattern Recognition*, and is/was a co-organizer of a series of Special Sessions/Workshops/Challenges in CVPR2020/FG2020/WACV2020/FG2019/BTAS2019, and so on.

Shiguang Shan (Senior Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He has been a Full Professor since 2010 and currently the Deputy Director of the CAS Key Laboratory of Intelligent Information Processing, CAS. He is a member of the CAS Center for Excellence in Brain Science and Intelligence Technology. He has published over 300 articles, with totally over 19 000 Google scholar citations. His research interests include computer vision, pattern recognition, and machine learning. He served as an Area Chair for many international conferences, including ICCV11, ICASSP14, ICPR12/14/19, ACCV12/16/18, FG13/18/20, BTAS18, and CVPR19/20. He was/is an Associate Editor of several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, *Neurocomputing*, CVIU, and PRL. He was a recipient of the China's State Natural Science Award in 2015 and the China's State S&T Progress Award in 2005 for his research work.

Xilin Chen (Fellow, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and over 300 articles in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a Fellow of ACM, IAPR, and CCF. He served as an organizing committee member for some conferences, including a General Co-Chair of FG13/FG18 and a Program Co-Chair of ICMI 2010. He is/was an Area Chair of CVPR 2017/2019/2020 and ICCV 2019. He is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, a Senior Editor of the *Journal of Visual Communication and Image Representation*, and an Associate Editor-in-Chief of the *Chinese Journal of Computers* and the *International Journal of Pattern Recognition and Artificial Intelligence* (Chinese).