




Adversarial Unsupervised Domain Adaptation Guided with Deep Clustering for Face Presentation Attack Detection

Yomna Safaa El-Din¹^a, Mohamed N. Moustafa²^b and Hani Mahdi¹^c

¹Computer and Systems Engineering Department, Ain Shams University, Cairo, Egypt

²Department of Computer Science and Engineering, The American University in Cairo, New Cairo, Egypt


Keywords: Biometrics, Face Presentation Attack Detection, Domain Adaptation, Deep Clustering, MobileNet.


Abstract: Face Presentation Attack Detection (PAD) has drawn increasing attentions to secure the face recognition systems that are widely used in many applications. Conventional face anti-spoofing methods have been proposed, assuming that testing is from the same domain used for training, and so cannot generalize well on unseen attack scenarios. The trained models tend to overfit to the acquisition sensors and attack types available in the training data. In light of this, we propose an end-to-end learning framework based on Domain Adaptation (DA) to improve PAD generalization capability. Labeled source-domain samples are used to train the feature extractor and classifier via cross-entropy loss, while unsupervised data from the target domain are utilized in adversarial DA approach causing the model to learn domain-invariant features. Using DA alone in face PAD fails to adapt well to target domain that is acquired in different conditions with different devices and attack types than the source domain. And so, in order to keep the intrinsic properties of the target domain, deep clustering of target samples is performed. Training and deep clustering are performed end-to-end, and experiments performed on several public benchmark datasets validate that our proposed Deep Clustering guided Unsupervised Domain Adaptation (DCDA) can learn more generalized information compared with the state-of-the-art classification error on the target domain.


1 INTRODUCTION

Face detection and recognition is an important topic in computer vision, it is used in many applications from which authentication is the most sensitive. Since the wide spread of smart mobile devices and the incorporation of latest vision technologies in these devices, end users find it more convenient to use their biometric data for authentication instead of classic passwords typing. On the other hand, this ease of use makes it easier for attacker to spoof the authentication system using pre-recorded biometric samples of the device user. Hence, the interest in developing reliable anti-spoofing or Presentation Attack Detection (PAD) techniques is increasing. Through the past years, several approaches were developed in literature (El-Din et al., 2020b) starting from basic methods relying on image processing and hand-engineered features, till approaches depending on automatically learnt features by deep-learning.

These approaches have succeeded to obtain perfect attack detection results on intra-dataset scenarios, where the dataset is split into training and testing subsets, so both subsets are coming from the same sensor model and acquisition environment. However, the main drawback of such methods is their lack of generalization to different environments and attack scenarios. The performance of the learnt representations in classifying the attack from the bona-fide (real) presentation degrades significantly when test data is captured by different sensor or in different settings or illumination conditions. In view of this, Domain Adaptation (DA) (Ganin et al., 2016) and Domain Generalization (DG) (Li et al., 2018c) were introduced recently in the PAD field. The target of DG is to learn representations that are robust across different domains, given samples from several source domains, such as in (Li et al., 2018a), (Shao et al., 2019), (Jia et al., 2020). While, DA aims at adapting a model trained on labeled source domain to a different target domain. Unsupervised DA (UDA) uses labeled samples from a source domain and unlabeled samples from a target domain, with a goal to achieve low clas-

^a <https://orcid.org/0000-0003-4959-4543>

^b <https://orcid.org/0000-0002-0017-3724>

^c <https://orcid.org/0000-0002-7442-9948>

sification error on the target domain though samples are unlabeled, by learning domain-invariant features.

For example, (Li et al., 2018b) experimented with both hand-crafted and deep learnt features in DA, however their approach was not end-to-end and the deep features did not generalize well. They achieved their best results using a combination of hand-crafted features. Adversarial training was used in DA for face PAD in (Wang et al., 2019) to learn an embedding space shared by both the source and target domain models. The training process is still not end-to-end where source pre-training, embedding adaptation and target classification are done separately.

In this paper, we focus on developing an end-to-end trainable solution for PAD based on DA, which focuses on improving the generalization of the model for cross-dataset testing without the need for several labeled source domains as in DG. Existing DA-based solutions solely aim to align the distribution of an unlabeled target domain to that of a different source domain, neglecting the specific nature of target domain. Target domain in face PAD is a different PAD dataset usually using a different device for authentication, in addition to different attack types in different illumination conditions. So solely trying to align the distribution of such different attacks scenarios to the distribution of attack scenarios in the labeled source dataset would not succeed, especially when the device used for authentication in one domain, is close to the one used for attack in the other domain, e.g. mobile device. So, we propose an approach that utilizes DA for PAD generalization to a different domain without neglecting the intrinsic properties of this target domain. We incorporate clustering based on deeply extracted features, for guiding the feature extraction network to generate features that are domain invariant, yet maintain the class-wise separability of the target dataset.

The main contributions of this work are: (1) proposing a novel end-to-end DA-based training architecture for the generalization of face PAD based; (2) utilize deep embedding clustering of target domain in guiding the DA process; (3) show substantial improvement on SOTA in cross-dataset evaluation on public benchmark face PAD datasets, with close to 0% cross-dataset error. The rest of the paper is organized as follows: Section 2 reviews the latest literature in face PAD and domain adaptation. Our proposed algorithm is explained in Section 3, followed by the experiments, benchmark datasets used and results in Section 4, then conclusions in Section 5.

2 RELATED WORK

2.1 CNN-based Face PAD

Recent software-based face presentation attack detection methods can be mainly categorized into texture-based and temporal-based techniques. The texture-based methods rely on extracting features from the frames that would identify if the presented image is fake or bona-fide. Features could be hand-crafted features as color texture (Boulkenafet et al., 2016), SIFT (Patel et al., 2016b) or SURF (Boulkenafet et al., 2017) which obtained good results in differentiating real from fake presentations. However, they are often sensitive to varying acquisition conditions, such as camera devices, lighting conditions and Presentation Attack Instruments (PAIs). Hence, the need to automatically learn and extract meaningful features directly from the data using deep representations, such as in (Nagpal and Dubey, 2018; El-Din et al., 2020b).

In addition to texture-based features, temporal-based models utilize the temporal information in face videos for better detection of attack presentations. Frame difference was combined with deep features in (Patel et al., 2016a). In (Feng et al., 2016) image quality information and motion information from optical flow were combined with neural network for classification. LSTM-CNN architecture was used in (Xu et al., 2015) and in (Wang et al., 2018) multiple RGB frames were used to estimate face depth information, and then two modules were used to extract short and long-term motion.

These methods obtain excellent results in intra-dataset testing, yet still fail to generalize to unseen environments and acquisition conditions. They show high cross-dataset evaluation errors, hence the need to incorporate domain adaptation techniques to decrease the discrepancy in distributions of the domain used for training and that used for deployment.

2.2 Unsupervised Domain Adaptation

Recently, Domain Adaptation (DA) has been introduced in computer vision, to tackle the problem of domain shift when applying models trained on a certain (source) domain to another (target) domain. Several methods, such as (Ganin et al., 2016), rely on adversarial training (Goodfellow et al., 2014) to guide the feature extraction module to generate domain-invariant features that make it harder for a domain discriminator to decide the original domain of the sample. Specifically, unsupervised DA uses labeled samples from the source domain in addition to unlabeled

samples from the target domain; to train a model that reduces the classification error on the unlabeled target domain.

Inspired by the success of DA in image classification (Pei et al., 2018), (Long et al., 2018), (Saito et al., 2018b), (Saito et al., 2018a), (Kurmi and Namboodiri, 2019), (Zhang et al., 2019), (Tang and Jia, 2020), (Kang et al., 2020), we believe that it can be used to address the problem of generalization in face PAD. A model fine-tuned on certain small-sized face PAD dataset fails to generalize when testing on different PAD domains with different domain. The learnt features become specific to the subjects or sensors available in the source dataset. Hence, by using domain adaptation in face PAD, the model will be guided to learn domain-invariant features that can differentiate between bona-fide and attack face videos regardless of the instance origin. However, learning domain invariant features can hurt classification of the target face PAD dataset by ignoring the fine-level class-wise structure of this target since the attack samples are generated with different instruments, and bona-fide samples may be captured by different sensors. Hence, we propose to incorporate deep clustering of target samples to constraint the model to keep the discriminative structure of both classes in the target dataset.

2.3 Deep Unsupervised Clustering

Deep learning is adopted in clustering of deep visual features since Deep Embedded Clustering (DEC) (Xie et al., 2016). Clustering aims at categorizing unlabeled data into groups (clusters). A DEC is a method that jointly learns feature representations and cluster assignments, where a neural network is first pre-trained by means of an autoencoder and then fine-tuned by jointly optimizing cluster centroids in output space and the underlying feature representation using Kullback-Leibler divergence minimization. Later, variants of DEC have emerged, such as (Guo et al., 2018) which adds data augmentation.

Unlike DEC, which require layer-wise pretraining as well as non-joint embedding and clustering learning, **DEE**P Embedded Regular**I**zed Clus**T**ering (DEPICT) (Dizaji et al., 2017) utilizes an end-to-end optimization for training all network layers simultaneously using the unified clustering and reconstruction loss functions. DEPICT consists of a multi-layer convolutional autoencoder followed by a multinomial logistic regression function. The clustering objective function uses relative entropy (KL divergence) minimization, regularized by a prior for the frequency of cluster assignments. An alternating strategy is then followed to optimize the objective by updating pa-

rameters and estimating cluster assignments. Reconstruction loss functions is employed in the autoencoder to prevent the deep embedding function from overfitting. A joint learning framework is introduced to minimize the unified clustering and reconstruction loss functions together and train all network layers simultaneously.

Recently, clustering has been introduced in several domain adaptation methods. (Wang et al., 2019) proposed a method to alleviate the effects of negative transfer in adversarial domain matching between source and target representations. They proposed to simultaneously learn tightly clustered target representations while encouraging that each cluster is assigned to a unique and different class from the source. In (Tang et al., 2020), structural domain similarity is assumed and the clustering solution is constrained using structural source regularization. By minimizing the KL divergence between predictive label distribution of the network and an introduced auxiliary one; replacing the auxiliary distribution with that formed by ground-truth labels of source data implements the structural source regularization via a simple strategy of joint network training.

2.4 DA in Face PAD

Domain Adaptation (DA) and Domain Generalization (DG) have been utilized recently to reduce the gap between the target domain and the source domain during face PAD. (Shao et al., 2019) focuses on improving the generalization ability of face PAD methods from the perspective of the domain generalization. Adversarial learning was proposed to train multiple feature extractors to learn a generalized feature space. They also incorporated an auxiliary face depth supervision to further enhance the generalization ability. Later, a Single-Side Domain Generalization framework was proposed in (SSDG) (Jia et al., 2020) that is end-to-end. They proposed to learn a generalized feature space, where the feature distribution of the real faces is compact while that of the fake ones is dispersed among domains but compact within each domain.

One of the first work exploring DA for face PAD is (Li et al., 2018b) were both hand-crafted features and deep neural network learned features are adopted and compared in DA. (Li et al., 2018b) found that the deep learning based methods may not generalize well under cross-database testing scenarios, and their best results were achieved using concatenated CoALBP and LPQ feature in HSV and YCbCr color space.

A 3D CNN architecture tailored for the spatial-temporal input is proposed by (Li et al., 2018a) for enhancing the generalization capability of the network.

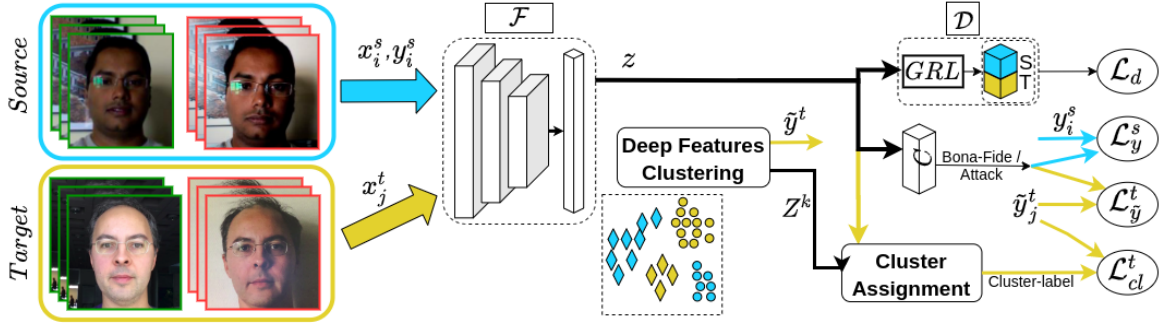


Figure 1: Architecture of the proposed Deep Clustering-guided-Domain Adaptation (DCDA) for face PAD. \mathcal{F} : Feature extraction network, \mathcal{D} : Domain Discriminator, GRL : Gradient Reverse Layer, \mathcal{C} : Categories Classifier, S : Source, T : Target. Bona-fide images are highlighted in green border, while attack images are highlighted in red. **Deep Features Clustering**: predicts target pseudo-labels \hat{y}^t and cluster centers Z^k . **Cluster Assignment**: assigns target features to clusters based on Student’s t -distribution.

A robust representation across different face spoofing domains is presented by introducing the generalization loss as the regularization term. Given training samples from several domains, the network is optimized such that the Maximum Mean Discrepancy (MMD) distances among different domains can be minimized. They performed the experiments by combining three publicly available face PAD datasets to create 10 protocols. In each protocol, data from one camera is set aside as the unseen target domain, and a subset of the remaining cameras are used as source domains.

ADA (Wang et al., 2019) is the first to incorporate adversarial domain adaptation in a learning approach to improve face PAD generalization capability. A source model optimized with triplet loss is first pre-trained in source domain, and then adversarial adaptation is used for training a target model to learn a shared embedding space by both the source and target domain models. Finally, target images are mapped with the target model to the embedding space and classified with k -nearest neighbors’ classifier. However, as the first attempt to use adversarial training for domain adaptation, the training is not performed end-to-end. In (Mohammadi et al., 2020), authors relied only on bona-fide samples of the target domain for DA. They hypothesize that, in a CNN trained for PAD given a source domain, some of the filters learned in the initial layers are robust filters that generalize well to the target dataset, whereas others are more specific to the source dataset. They propose to prune such filters that do not generalize well from one dataset to another in order to improve the performance of the network on the target dataset. Feature Divergence Measure (FDM) is computed to quantify the level of domain shift at a given layer in a CNN.

(Wang et al., 2020) proposed disentangled representation learning for cross-domain face PAD. Their

approach consists of Disentangled Representation learning (DR-Net) and Multi-Domain feature learning (MD-Net). DR-Net learns a pair of encoders via generative models that can disentangle PAD informative features from subject discriminative features. The disentangled features from different domains are fed to MD-Net which learns domain-independent features for the final cross-domain face PAD task. They tested single-source to single-target cross-domain PAD and also multi-source to multi-target and obtained state of the art results on four public datasets. Their later work (DR-UDA) (Wang et al., 2021) consists of three modules, ML-Net, UDA-Net and DR-Net. ML-Net uses the labeled source domain face images to learn a discriminative feature representation. 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. Furthermore, DR-Net disentangles the features irrelevant to specific domains by reconstructing the source and target domain face images from the common feature space.

3 METHODOLOGY

In this section, we introduce the frameworks of unsupervised DA and unsupervised clustering. Then, we present our proposed model for UDA in face PAD. Figure 1 shows a brief overview of the proposed architecture.

Since the most common target platform is mobile devices, we follow (El-Din et al., 2020a) and use latest architecture of MobileNet; MobileNetV3 (Howard et al., 2019) instead of the commonly used Resnet-50 (He et al., 2016). MobileNet is tuned for mobile phone CPUs which helps preserve the mobile battery life by reducing power consumption. With $\sim 80\%$

less parameters, MobileNetV3 achieves comparable ImageNet accuracy as Resnet50 with reduced inference time.

3.1 Deep Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA), depends on having a set of labeled source samples $S = \{(x_i, y_i)\}_{i=1}^{N_s}$ and another set of unlabeled samples from target domain $T = \{x_j\}_{j=1}^{N_t}$. The goal is to train a model that is capable of achieving low classification errors on the unlabeled target domain guided by the labeled source samples. The feature extraction module is trained to be able to extract features that benefit the categories classification without differentiating the domain origin of the sample.

As (DANN) (Ganin et al., 2016), adversarial training is incorporated to guide the feature extraction module, \mathcal{F} , to generate features that confuse a domain discriminator, \mathcal{D} , to not be able to determine the domain of the input features. The categories (task) classifier, \mathcal{C} , is then trained on top of these generated domain-invariant features; using the labeled source samples, to decide the final classification label.

The task classification loss is calculated as

$$L_y^s = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{L}_y(\mathcal{C}(\mathcal{F}(x_i)), y_i), \quad (1)$$

where \mathcal{L}_y is categorical cross-entropy loss, \mathcal{F} is the feature extractor network and L_y^s is the task classification loss from all source samples using. Similarly, domain discrimination loss,

$$L_d = \frac{1}{N_s + N_t} \sum_{m=1}^{(N_s + N_t)} \mathcal{L}_d(\mathcal{D}(\mathcal{F}(x_m)), d_m), \quad (2)$$

where \mathcal{L}_d is categorical cross-entropy loss, d_m is domain label, zero for source samples, and one otherwise. This loss is minimized over the parameters of ffd while maximized over the parameters of \mathcal{F} via the gradient reverse layer (GRL).

3.2 Proposed DC-guided UDA for Face PAD

For handling the problem of generalization in face PAD, we propose to use UDA, in combination with Deep Embedding Clustering (DEC) of the unlabeled target samples during training. Motivation for UDA is to alleviate the shift between the source and target domains. However, we do not want to lose the target properties for each class.

Aligning both source and target domains in face PAD with source and target coming from different sensors and attack instruments, might lead to target samples being misclassified and shifted towards the wrong class. For example, a target mobile attack instance can be assigned to the closest source sample which might be bona-fide class if bona-fide samples of source dataset are captured with same instrument (mobile device). So motivation for adding target clustering is to preserve the class-wise separation of target domain samples. Which together with adversarial DA, will guide \mathcal{F} to generate features that reduce domain shift without corrupting the class-wise separability of target domain.

Algorithm 1: Training of DCDA: Deep Clustering-guided-Domain adaptation for face PAD.

Let $\{\theta_{\mathcal{F}}, \theta_{\mathcal{C}}, \theta_{\mathcal{D}}\}$ be the learnable parameters for each model component.

Let $\{Z_{BF}^k, Z_A^k\}$ be the learnable cluster centers for bona-fide and attack classes respectively.

Input:

Labeled source videos $S : (X^s, Y^s)$ and unlabeled target videos $T : (X^t)$

Batch size: B

Output:

Feature extractor: $\mathcal{F}(\cdot)$

Classifier: $\mathcal{C}(\cdot)$

Deep Discriminative Clustering:

Fix model parameters

$\{z_i^s\} = \mathcal{F}(x_i^s)$ for all $x_i^s \in X^s$

$\{z_j^t\} = \mathcal{F}(x_j^t)$ for all $x_j^t \in X^t$

$Z_{cs}^k = \text{avg}(\{z_i^s\})$ for $y_i^s = c \forall c \in \{BF, A\}$

$\tilde{y}^t, Z_{BF^t}^k, Z_{A^t}^k \leftarrow$ k-means clustering of $\{z_j^t\}$ using

$Z_{BF^s}^k, Z_{A^s}^k$ as initial centers

for $c \in \{BF, A\}$ **do**

$Z_{cs}^k = \text{avg}(\{z_i^s\})$ for $y_i^s = c$

$Z_c^k = \text{avg}(Z_{cs}^k, Z_{ct}^k)$

$ep = 0$

while $ep < \text{max_epochs}$ **do**

for $b = 0$ **to** iter_per_epoch **do**

Draw random batch $\{(x_i^s, y_i^s)\}_{i=1}^B,$

$\{(x_j^t, \tilde{y}_j^t)\}_{j=1}^B$

$\theta_{\mathcal{F}} = \theta_{\mathcal{F}} - \nabla_{\theta_{\mathcal{F}}} (L_y^s + L_y^t + L_d + L_{cl}^t)$

$\theta_{\mathcal{C}} = \theta_{\mathcal{C}} - \nabla_{\theta_{\mathcal{C}}} (L_y^s + L_y^t)$

$\theta_{\mathcal{D}} = \theta_{\mathcal{D}} - \nabla_{\theta_{\mathcal{D}}} L_d$

$Z_c^k = Z_c^k - \nabla_{Z_c^k} L_{cl}^t, \forall c \in \{BF, A\}$

end for

Update target pseudo-labels \tilde{y}^t based on $\{z_j^t\}$ distance to Z_{BF}^k and Z_A^k

$ep = ep + 1$

end while

3.2.1 Deep Clustering for DA

Our training follows the unsupervised deep clustering methods (Xie et al., 2016), (Dizaji et al., 2017) which alternates between cluster assignment while fixing model parameters, then model update while fixing these cluster assignment. At the start of each epoch, k-means clustering is performed on the deep features generated by \mathcal{F} to generate pseudo-labels, \tilde{Y}^t , for the unlabeled target samples. Then, during epoch iterations, two losses based on Kullback-Leibler (KL) divergence (Xie et al., 2016) are minimized to update the parameters of \mathcal{F} , C and cluster centroids Z^k via back-propagation.

These learnable centroids $Z^k = \{Z_{BF}^k, Z_A^k\}$ for each of the bona-fide and attack classes are re-updated at the start of each epoch, while fixing the model parameters. Guided by the labels of source samples, and the source features generated by the current \mathcal{F} , clusters centers for the source domain; Z_{cs}^k , can be obtained in the embedding space. On the other hand, for the unlabeled target samples, k-means clustering is used on the generated latent features of all target samples. This obtains both pseudo-labels for all target instances in training, \tilde{Y}^t , and clusters centers for the target domain, Z_{ct}^k . Finally, the learnable cluster center for each class Z_c^k is updated to be the mean of both Z_{ct}^k and Z_{cs}^k .

During training iterations of an epoch, target samples are used to minimize KL divergence two-way. The loss to be minimized can be written as

$$L_{dec} = KL(Q||P) + L_{reg} \quad (3)$$

$$= \frac{1}{N} \sum_{j=1}^N \sum_{k=1}^K q_{jk} \log \frac{q_{jk}}{p_{jk}} + \sum_{k=1}^K \hat{q}_k \log \hat{q}_k,$$

where P^t is the cluster assignments for target samples and Q^t is an auxiliary target distributions, and the purpose of KL divergence minimization is to decrease the distance between the model predicted P^t and the distribution Q^t . The second term follows (Krause et al., 2010) for incorporating class balance to avoid degenerate solutions, where $\hat{q}_k = \frac{1}{N_t} \sum_{j=1}^{N_t} q_{jk}^t$.

As in (Dizaji et al., 2017), optimization of loss in equation 3 alternates between updating auxiliary distribution Q^t then using Q^t to update model parameters. Q^t is calculated in closed-form solutions as

$$q_{jk}^t = \frac{p_{jk}^t / (\sum_{j'} p_{j'k}^t)^{\frac{1}{2}}}{\sum_{k'} p_{jk'}^t / (\sum_{j'} p_{j'k'}^t)^{\frac{1}{2}}}. \quad (4)$$

For further regulation of target clustering, we use the previously estimated target pseudo-labels as part of Q^t by setting $q_j^t = 0.5 * q_j^t + 0.5 * \tilde{y}_j^t$.

Then using calculated P^t and Q^t , parameters of \mathcal{F} and C are updated by minimizing

$$L_{cl}^t = -\frac{1}{N_t} \sum_{j=1}^{N_t} \sum_{k=1}^K q_{jk}^t \log p_{jk}^t, \quad (5)$$

As mentioned earlier, we use KL divergence minimization with target domain samples for two losses which update parameters of feature extraction module \mathcal{F} via backpropagation. The first loss additionally aims to update the classifier C as well, and the second loss updates the cluster centroids Z^k . For the first loss ($L_{\tilde{y}}^t$), we set P^t as the classifier prediction probabilities after softmax; $p_j^t = \text{softmax}(C(\mathcal{F}(x_j^t)))$, so that it becomes like cross-entropy classification loss using pseudo-labeled target samples.

For the second loss (L_{cl}^t), P^t is estimated using the Student's t -distribution to measure the similarity between target features Z^t and cluster centroids Z^k as in (Xie et al., 2016)

$$p_{jc}^t = \frac{(1 + \|z_j^t - Z_c^k\|^2 / \alpha)^{-\frac{\alpha+1}{2}}}{\sum_{c'} (1 + \|z_j^t - Z_{c'}^k\|^2 / \alpha)^{-\frac{\alpha+1}{2}}}.$$

Finally, the estimated pseudo-labels for target samples are used to update the parameters of both the feature extractor \mathcal{F} and the classifier C by minimizing the following task classification loss

$$L_{\tilde{y}}^t = \frac{1}{N_t} \sum_{j=1}^{N_t} \mathcal{L}_{\tilde{y}}(C(\mathcal{F}(x_j)), \tilde{y}_j), \quad (6)$$

where $\mathcal{L}_{\tilde{y}}$ is categorical cross-entropy loss.

3.2.2 Complete Model Learning

The complete end-to-end training methodology of our proposed DC-guided-DA for face PAD is listed in Algorithm 1. We use only one frame per video.

4 EXPERIMENTS AND RESULTS

4.1 Face PAD Datasets

Table 1 summarizes the total number of samples present in each subset of the datasets used, in addition to the Presentation Attack Instruments (PAI) used and the sensors used in recording videos for authentication.

Replay-Attack (Chingovska et al., 2012) is one of the earliest datasets presented in literature for the problem of face spoofing. It consists of 1200 short videos from 50 different subjects with resolution 320×240 from 50 different subjects. Attack scenario include

Table 1: Number of samples per class per subset for each used PAD dataset.

Database	PAI	Sensor used for authentication	Subset	Bona-fide	Attack	Total
Replay-Attack	1) PR (A4)	(1) Webcam in MacBook laptop	train	300	60	360
	2) VR on iPhone		devel	300	60	360
	3) VR on iPad		test	400	80	480
MSU-MFSD	1) PR (A3)	1) Webcam in MacBook Air 2) FC of Google Nexus5 Mob	train	90	30	120
	2) high-def VR on iPad 3) VR on iPhone		test	120	84	204
Replay-Mobile	1) PR (A4)	1) FC of iPad Mini2 Tablet 2) FC of LG-G4 Mobile	train	192	120	312
	2) VR on matte-screen		devel	256	160	416
			test	192	110	302

FC: Front-Camera, **PR**: Hard-copy print of high-res photo, **VR**: Video replay

Table 2: Results of Proposed DC-guided-DA for Face-PAD in ACER% at threshold 0.5.

train→test	RA→M	RA→RM	M→RA	M→RM	RM→RA	RM→M	Average
Source-only	34	49.8	39.4	15.6	42.3	42	37.18
DA w/o clustering	29.6	47.2	49.25	11.35	45	2.9	30.88
DCDA w/o L_y^2	18.35	49.2	10.40	2.25	11.65	37.80	19.94
DCDA	0	0	0.15	1.6	1.15	1.65	0.76

RA: Replay-Attack, **M**: MSU-MFSD, **RM**: Replay-Mobile

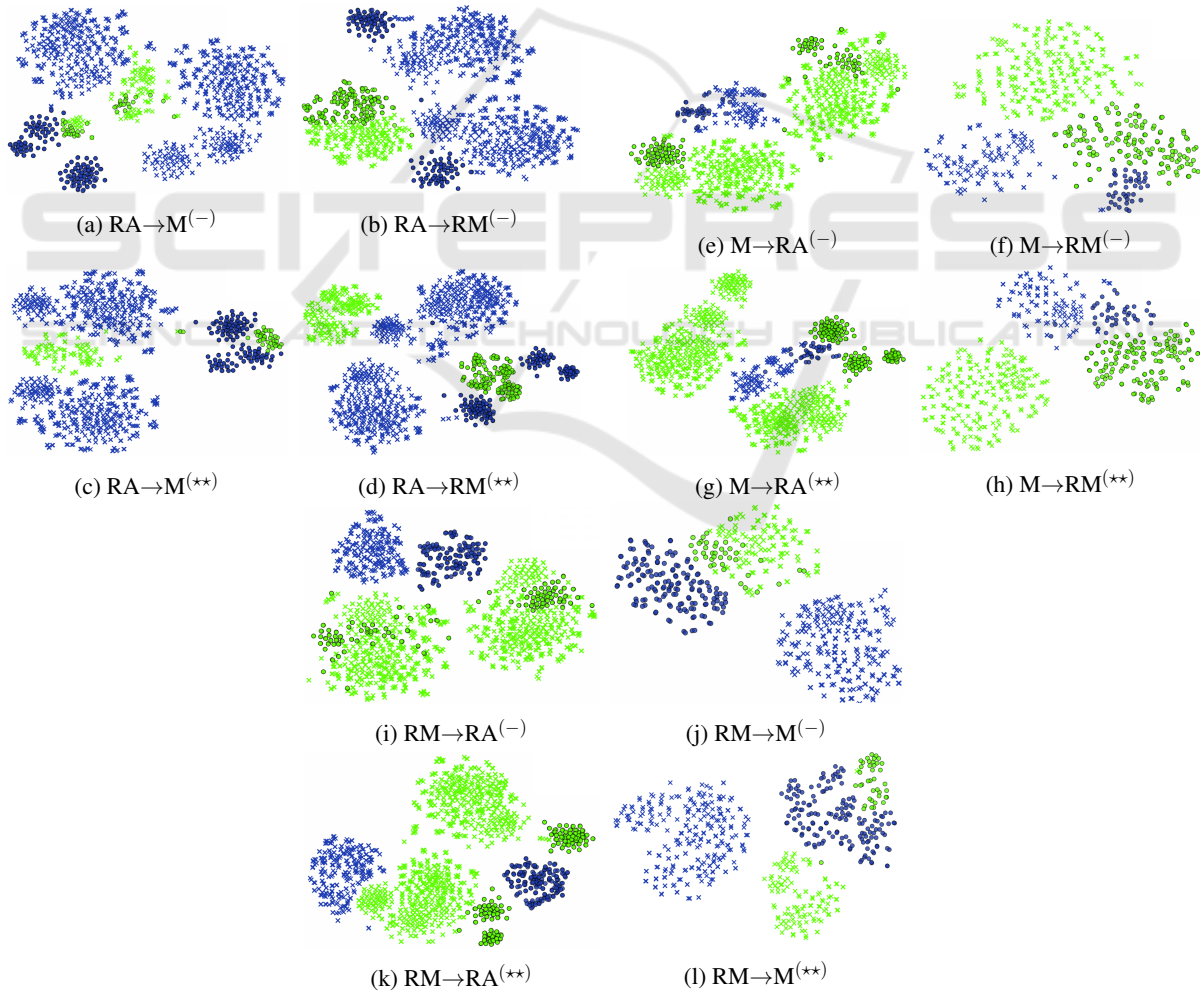


Figure 2: t-SNE visualization analysis. Upper row $(-)$: DA without clustering, Bottom row $(**)$: Proposed DC-guided-DA. *Blue*: Source, *Green*: Target, *o*: Bona-fide, *x*: Attack. Best viewed in color.

Table 3: Comparison with SOTA in HTER%.

	RA→M	M→RA	Average
KSA [§] (Li et al., 2018b)	18.6*	23.3*	20.95
ADA (Wang et al., 2019)	30.5	5.1	17.8
PAD-GAN (Wang et al., 2020)	23.2	8.7	15.95
SSDG (Jia et al., 2020)	7.38**	11.7**	9.54
DCDA (Proposed)	0	0.15	0.08

* On concatenated CoALBP and LPQ features in HSV and YCbCr color space

** Source-domain includes two other datasets

”hard-copy print-attack”, ”mobile-photo attack” and ”high-definition screen attack”. Attacks are presented to the sensor (regular webcam) either with a ”fixed” tripod, or by an attacker holding the presenting device (printed paper or replay device) with his/her ”hand”.

MSU Mobile Face Spoofing Database (MSU-MFSD) (Wen et al., 2015) targets the problem of face spoofing on smartphones . The dataset includes real and spoofed videos from 35 subjects . Two devices were used, the webcam of a MacBook Air with resolution 640×480 and the front facing camera of a smartphone with 720×480 resolution. Three attack scenarios are used: print-attack on A3 paper, video replay attack on the screen of an iPad and video replay attack on a smartphone.

Replay-Mobile (Costa-Pazo et al., 2016) was released by the same research institute that released Replay-Attack. It has 1200 short videos from 40 subjects captured by two mobile devices at resolution 720×1280 . Each subject has ten bona-fide accesses and 16 attack videos under different attack modes. Two types of attack are present: photo-print and matte-screen attack displaying digital-photo or video.

4.2 Experimental Setup

Our experiments were performed on NVIDIA GeForce 840m GPU with CUDA version 11.0. Bob package (Anjos et al., 2012) was used for datasets management and PyTorch was used for models and training. Evaluation metrics for PAD are the ISO/IEC 30107-3:2017¹ metrics. Attack Presentation Classification Error Rate (APCER), Bona-fide Presentation Classification Error Rate (BPCER) and their Average Classification Error Rate (ACER) ($(APCER + BPCER)/2$) is used for reporting results in the tables.

4.3 Results and Discussion

Table 2 presents results of our proposed DC-guided UDA for face PAD on the 3 benchmark face datasets used. Results are reported as the average ACER %

¹<https://www.iso.org/standard/67381.html>

of three runs, ACER is calculated on the test subset of the target dataset. The first row represents the results obtained by fine-tuning a MobileNetV3 classification network on source dataset only without domain adaptation. We performed experiments to study the influence of each model component on the overall performance of the algorithm. Clustering components and losses were removed and only Domain Adaptation was performed, results in the second row of Table 2 show only slight improvement over source-only trained models. Then, adding clustering components with target pseudo-labels estimation and target clustering loss L_{cl}^t , but without updating the classifier C with target classification loss $L_{\tilde{y}}^t$, yielded a significant decrease in the target classification error on most datasets as shown in third row. However, though feature extraction network is trying to learn domain-invariant features, the classifier trained on source-samples only still fails in some cases to achieve low errors on some target datasets. For example, the classifier trained on Replay-Attack dataset fails to discriminate the attack and bona-fide samples on Replay-Mobile dataset.

Finally, the last row shows results obtained by our full proposed DCDA framework, which achieves near-perfect classification of the unlabeled target samples. Comparison with state-of-the art DA-based face PAD solutions is provided in Table 3 showing superiority of our proposed DC-guided-DA framework. Furthermore, t -SNE visualization analysis is presented in Figure 2, comparing our proposed architecture, with models trained using Domain Adaptation only. The visualizations show that our proposed framework could align the classification boundaries for both source and target datasets, it also shows the diversity of attack and sensors types present in the same dataset that form clusters in the same class of the same dataset, for example Replay-Attack in Figure 2 parts 2c, 2d, 2g and 2k.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed an approach that exploits unsupervised adversarial domain adaptation guided with target clustering, in order to improve the generalization ability for face PAD. Specifically, our framework utilizes UDA to learn domain invariant features that could leverage from the labeled source samples to classify the unlabeled samples from target domain. Yet, the approach succeeds to preserve the intrinsic properties of the target domain via deep clustering of target embedding features. Our approach is trained in an end-to-end fashion and succeeds to reach perfect adaptation to the target domain when evaluated on public benchmark datasets, reaching only 0 - 2% cross-dataset error. Our future work would focus on evaluating on more variable datasets, in addition to reducing the dependency of the model during training on target domain samples from both classes, trying to let the model focuses on learning from bona-fide samples with minimal attack samples contribution.

REFERENCES

- Anjos, A., Shafey, L. E., Wallace, R., Günther, M., McCool, C., and Marcel, S. (2012). Bob: a free signal processing and machine learning toolbox for researchers. In *20th ACM Conference on Multimedia Systems (ACMMM)*, Nara, Japan.
- Boulkenafet, Z., Komulainen, J., and Hadid, A. (2016). Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830.
- Boulkenafet, Z., Komulainen, J., and Hadid, A. (2017). Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145.
- Chingovska, I., Anjos, A., and Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7.
- Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., and Marcel, S. (2016). The replay-mobile face presentation-attack database. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7.
- Dizaji, K. G., Herandi, A., Deng, C., Cai, W., and Huang, H. (2017). Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *IEEE international conference on Computer Vision*, pages 5747–5756.
- El-Din, Y. S., Moustaf, M. N., and Mahdi, H. (2020a). On the effectiveness of adversarial unsupervised domain adaptation for iris presentation attack detection in mobile devices. In *ICMV'20*.
- El-Din, Y. S., Moustafa, M. N., and Mahdi, H. (2020b). Deep convolutional neural networks for face and iris presentation attack detection: survey and case study. *IET Biometrics*, 9:179–193(14).
- Feng, L., Po, L.-M., Li, Y., Xu, X., Yuan, F., Cheung, T. C.-H., and Cheung, K.-W. (2016). Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *Journal of Visual Communication and Image Representation*, 38:451 – 460.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Lavolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.
- Guo, X., Zhu, E., Liu, X., and Yin, J. (2018). Deep embedded clustering with data augmentation. volume 95 of *Proceedings of Machine Learning Research*, pages 550–565. PMLR.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., and Le, Q. (2019). Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324.
- Jia, Y., Zhang, J., Shan, S., and Chen, X. (2020). Single-side domain generalization for face anti-spoofing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kang, G., Jiang, L., Wei, Y., Yang, Y., and Hauptmann, A. G. (2020). Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*.
- Krause, A., Perona, P., and Gomes, R. G. (2010). Discriminative clustering by regularized information maximization. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 775–783. Curran Associates, Inc.
- Kurmi, V. K. and Nambodiri, V. P. (2019). Looking back at labels: A class based domain adaptation technique. In *International Joint Conference on Neural Networks (IJCNN)*.
- Li, H., He, P., Wang, S., Rocha, A., Jiang, X., and Kot, A. C. (2018a). Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652.

- Li, H., Li, W., Cao, H., Wang, S., Huang, F., and Kot, A. C. (2018b). Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809.
- Li, H., Pan, S. J., Wang, S., and Kot, A. C. (2018c). Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.
- Long, M., CAO, Z., Wang, J., and Jordan, M. I. (2018). Conditional Adversarial Domain Adaptation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 1640–1650. Curran Associates, Inc.
- Mohammadi, A., Bhattacharjee, S., and Marcel, S. (2020). Domain adaptation for generalization of face presentation attack detection in mobile settings with minimal information. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1001–1005.
- Nagpal, C. and Dubey, S. R. (2018). A performance evaluation of convolutional neural networks for face anti spoofing. *CoRR*.
- Patel, K., Han, H., and Jain, A. K. (2016a). Cross-database face antispoofing with robust feature representation. In You, Z., Zhou, J., Wang, Y., Sun, Z., Shan, S., Zheng, W., Feng, J., and Zhao, Q., editors, *Biometric Recognition*, pages 611–619, Cham. Springer International Publishing.
- Patel, K., Han, H., and Jain, A. K. (2016b). Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283.
- Pei, Z., Cao, Z., Long, M., and Wang, J. (2018). Multi-adversarial domain adaptation.
- Saito, K., Ushiku, Y., Harada, T., and Saenko, K. (2018a). Adversarial dropout regularization. In *International Conference on Learning Representations*.
- Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (2018b). Maximum classifier discrepancy for unsupervised domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732.
- Shao, R., Lan, X., Li, J., and Yuen, P. C. (2019). Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10015–10023.
- Tang, H., Chen, K., and Jia, K. (2020). Unsupervised domain adaptation via structurally regularized deep clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tang, H. and Jia, K. (2020). Discriminative adversarial domain adaptation. *ArXiv*, abs/1911.12036.
- Wang, G., Han, H., Shan, S., and Chen, X. (2019). Improving cross-database face presentation attack detection via adversarial domain adaptation. In *2019 International Conference on Biometrics (ICB)*, pages 1–8.
- Wang, G., Han, H., Shan, S., and Chen, X. (2020). Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6677–6686.
- Wang, G., Han, H., Shan, S., and Chen, X. (2021). Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:56–69.
- Wang, R., Wang, G., and Henao, R. (2019). Discriminative clustering for robust unsupervised domain adaptation. *ArXiv*, abs/1905.13331.
- Wang, Z., Zhao, C., Qin, Y., Zhou, Q., and Lei, Z. (2018). Exploiting temporal and depth information for multi-frame face anti-spoofing. *CoRR*.
- Wen, D., Han, H., and Jain, A. K. (2015). Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761.
- Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In Balcan, M. F. and Weinberger, K. Q., editors, *International conference on machine learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA. PMLR.
- Xu, Z., Li, S., and Deng, W. (2015). Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 141–145.
- Zhang, Y., Tang, H., Jia, K., and Tan, M. (2019). Domain-symmetric networks for adversarial domain adaptation.