IMPROVING CROSS-DATASET PERFORMANCE OF FACE PRESENTATION ATTACK DETECTION SYSTEMS USING FACE RECOGNITION DATASETS

Amir Mohammadi, Sushil Bhattacharjee, and Sébastien Marcel

Idiap Research Institute, Switzerland

ABSTRACT

Presentation attack detection (PAD) is now considered critically important for any face-recognition (FR) based access-control system. Current deep-learning based PAD systems show excellent performance when they are tested in intra-dataset scenarios. Under cross-dataset evaluation the performance of these PAD systems drops significantly. This lack of generalization is attributed to *domain-shift*. Here, we propose a novel PAD method that leverages the large variability present in FR datasets to induce invariance to factors that cause domain-shift. Evaluation of the proposed method on several datasets, including datasets collected using mobile devices, shows performance improvements in cross-dataset evaluations.[1]

***Index Terms*—** mobile biometrics, presentation attack detection, cross-dataset evaluation, domain generalization

1. INTRODUCTION

The past couple of years have seen a surge in the use of face-recognition (FR) technology on mobile platforms. This uptake has been driven by mainly two factors, the extremely high recognition accuracy achieved by modern, deep learning based face recognition (FR) systems [1, 2, 3], and the convenience of using FR over to other biometrics modalities. Nonetheless, state-of-the-art FR systems remain highly vulnerable to *presentation attacks* (PA, also referred to as spoof-attacks) [4]. In this work we consider two kinds of PAs:

1. *Print* attack: where the attacker presents a printed photograph of the intended victim to the camera of the FR system under attack, and

2. *Replay* attack: where the biometric sensor (camera of the FR system) is presented with a video of the intended victim being replayed on a digital display such as the display of a tablet computer. In this context, *bona fide* (BF) sample [5] refers to non-attack presentations. Countermeasures against PAs are called Presentation attack detection (PAD) methods.

Various research groups working on PAD have publicly shared face-PAD datasets [6, 7, 8], several collected using mobile-devices. These datasets include protocols defining mutually disjoint data subsets for *training* and *evaluation* of face-PAD methods. After training using a dataset (the *source* dataset), two scenarios are possible for evaluating a face-PAD system:

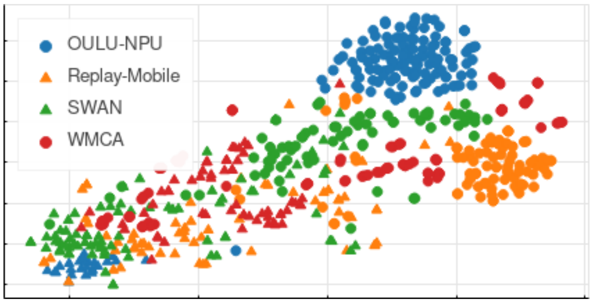

Fig. 1: t-SNE [14] plot of the embeddings of a CNN-based PAD system (DeepPixBiS [9]) for four datasets. Samples with the same color belong to the same face-PAD dataset. Triangles are BF samples and circles are PA samples. We observe that, within each class, samples from each dataset are clustered. This may be attributed to the *domain shift* present between face-PAD datasets.

1. *intra-dataset* evaluation: the evaluation set is taken from the source dataset, or,
2. *cross-dataset* evaluation: the evaluation set is taken from a different dataset (not the source dataset).

Current convolutional neural networks (CNN) based face PAD systems perform well in intra-dataset evaluation scenarios [9, 10, 11]. Typically, however, their performance degrades significantly when tested in cross-dataset scenarios [9, 10, 11]. Such generalization issues in machine learning models have been attributed to *domain shift* (also called covariate shift, or dataset bias) present between two datasets [12, 13].

The problem of domain-shift is illustrated in Figure 1. The figure shows a t-distributed stochastic neighbor embedding (t-SNE) plot [14] where feature-vectors extracted using a certain face-PAD CNN have been projected onto two dimensions using a specific multi-dimensional scaling method. Feature-vectors for BF (triangles) and PA samples (circles) in four different datasets (each identified by a different color) are shown in this plot. Considering only the BF samples, we note that samples from different datasets form distinct clusters. That is, BF samples of different datasets produce feature-vectors with different distributions. (Similar observations can be made for the PA samples of the different datasets as well.) This exemplifies the problem of domain-shift.

In face-PAD datasets, domain-shift may be caused by a variety of factors, including: the camera device, resolution of images, distance of the subject to the camera, the instrument used to create the attack, lighting conditions, and identity. These *factors* are *nuisance* factors to a face PAD system. Ideally a PAD system should be invariant to these factors when classifying face images.

Several methods for inducing invariance to nuisance factors in the learning process [15, 16, 17, 18, 19] have been proposed. How-

This work was funded by the Research Council of Norway (Grant No. IKTPLUSS 248030/O70). Amir Mohammadi (amir.mohammadi@idiap.ch) is also affiliated with École Polytechnique Fédérale de Lausanne (EPFL).

[1] Source code: https://gitlab.idiap.ch/bob/bob.paper.icassp2020_facepad_generalization_infovae

ever, most works induce invariance to factors that are known *a priori* and are explicitly labeled. Identifying all the nuisance factors in a given dataset and labeling them can prove difficult [20]. For example, one of the nuisance factors in face PAD corresponds to the lighting conditions but categorizing images with respect to lighting conditions is a tedious, labor intensive, and subjective process.

In this work, we hypothesize that all factors present in an FR dataset (which contains only BF samples) are nuisance factors in a face PAD system. Among others, these include factors such as age, pose, illumination conditions, and facial-makeup. By explicitly modeling these factors in an unsupervised manner, we aim to induce invariance to these factors in a face PAD system. While these factors also exist in face PAD datasets, only a small variety of each factor is represented in a face PAD dataset. For example, face PAD datasets are usually collected with less than 10 camera devices, for 50 to 150 identities, and have limited variations in lighting conditions. FR datasets, on the other hand, contain millions of face images with hundreds of thousands of identities which are adequately varied [21]. Being such large and varied datasets, they can be used to adequately model some nuisance factors of face-PAD systems.

To the best of our knowledge, this is the first work to take advantage of FR datasets to improve generalization of face-PAD systems in an *unsupervised* manner. Other works such as [10, 22] use multi-task learning of both FR and PAD which requires the face images to be labeled according to identities in both FR and PAD datasets. However, identity labels are not strictly necessary for PAD, and most PAD datasets do not include identity labels. Moreover, these methods only use an FR dataset for initialization of a multi-task network. That is, the FR part of the network is first trained on a large FR dataset. The network is subsequently trained on a smaller PAD dataset with a few identities for both tasks of PAD and FR.

Previous works related to the proposed method are presented in Section 2. The proposed method is detailed in Section 3. Implementation details are outlined in Section 4. Experiments are described in Section 5 and conclusions are made in Section 6.

2. RELATED WORK

One recent method that proposes to induce invariance to all nuisance factors in an unsupervised manner is *unsupervised adversarial invariance* (UAI) [19]. In UAI, a neural network is trained simultaneously for two tasks: the required primary task (classification or regression), as well as reconstruction of the input. After a few initial layers, the network splits into two branches, each dedicated to optimizing one task. The initial layers produce two embeddings: e_1 and e_2 which will be used in the two task-specific branches. The reconstruction branch takes two inputs: e_2, and $\hat{e}_1$, a noisy version of e_1. The input to the branch responsible for the primary task is only e_1. Two adversarial losses are added which make sure e_1 and e_2 do not contain duplicate information (see [19]). For reconstruction, most factors of the data are needed to correctly reconstruct the input. Since reconstruction is done using e_2 and $\hat{e}_1$ (which is noisy), it is assumed that most factors of the data will be represented by e_2 to guarantee correct reconstruction of input. Also, since by construction e_1 and e_2 do not contain duplicate information, only factors crucial for the primary task will be represented by e_1. However, in this approach, if the dataset contains a bias that may simplify the primary task, there is no guarantee that it will not be represented in e_1. In fact, e_1 could include the bias of the dataset. Moreover, in this method, the nuisance factors are modeled using the dataset for the primary task. In case of face PAD, these datasets may not fully represent all the possible nuisance factors. In [23], where the authors use UAI for face PAD, no cross-dataset performance evaluation is reported.

The inter-session variability (ISV) technique proposed in [24, 25] explicitly models within-class variations (nuisance factors) in Gaussian Mixture Model (GMM)-based biometric recognition systems [26, 27]. Assuming that samples from all classes (identities in FR) have the same nuisance factors and that these factors are contained in a linear subspace of GMM mean supervector space, a training mechanism is proposed to explicitly model these nuisance factors in the GMM mean supervector space. Once these factors are modeled, given a face image and its GMM-based mean supervector, its nuisance factors are estimated and their effect is removed from the mean supervector. This obtained mean supervector is used for classification instead of the original mean supervector. While this method has been successfully applied on FR, it is limited to GMM-based systems which use hand-crafted features. Since then, many deep learning based FR algorithms have outperformed GMM-based methods [1, 2, 3].

3. PROPOSED METHOD

Many nuisance factors can cause domain shifts in face PAD datasets. In this work, we assume that all nuisance factors present in BF face images are also present in PA face images. For example, factors such as identities, lighting conditions, and camera devices can be different in both BF and PA samples between datasets. Here, we propose a method to explicitly model these common nuisance factors using an FR dataset. We assume that these factors are well represented in an FR dataset which contains millions of BF face images. By explicitly modeling these factors, we can induce invariance to these factors in a PAD system.

More specifically, assume that each face image $\mathbf{I}$ is generated through a function f and some noise, ϵ:

$$\mathbf{I} = \mathrm{f}(\mathbf{y}, \mathbf{z_1}, \mathbf{z_2}) + \epsilon \tag{1}$$

where $\mathbf{y}$ is the variable that we want to predict – whether $\mathbf{I}$ is a PA, $\mathbf{z_1}$ and $\mathbf{z_2}$ are multivariate latent variables. The variable $\mathbf{z_1}$ represents all the nuisance factors present in BF samples that are present in PA samples as well, whereas $\mathbf{z_2}$ represents all other nuisance factors that are exclusive to PAs. The variable $\mathbf{z_1}$ may encapsulate information about gender, pose, identities, lighting condition, camera characteristics, and so on. The variable $\mathbf{z_2}$ may contain information about the presentation attack instrument (PAI). In this work, we will not model $\mathbf{z_2}$ or try to induce invariance to $\mathbf{z_2}$. However, if some factors present in $\mathbf{z_2}$ are known and labeled, it is possible to induce invariance to these factors using traditional invariance induction methods such as [16].

For simplicity, assume that f can be modeled as the sum of two other functions:

$$\mathrm{f}(\mathbf{y}, \mathbf{z_1}, \mathbf{z_2}) = \mathrm{g}(\mathbf{z_1}) + \mathrm{h}(\mathbf{y}, \mathbf{z_2}) \tag{2}$$

where the functions g and h each produce an image given their respective latent variables as input. The final image is the sum of these two images plus some noise. Given an image $\mathbf{I}$, we assume that $\mathbf{z_1}$ may be estimated through some function, e:

$$\mathbf{z_1} = \mathrm{e}(\mathbf{I}) \tag{3}$$

and that the function g is also known. This allows us to reconstruct a face image using only $\mathbf{z_1}$:

$$\mathbf{I}_{\mathbf{z_1}} = \mathrm{g}(\mathbf{z_1}) = \mathrm{g}(\mathrm{e}(\mathbf{I})) \tag{4}$$

Then, given a BF or PA face image, we can approximate the output of h as:

$$\mathbf{I}_{\mathbf{y},\mathbf{z_2}} = \mathrm{h}(\mathbf{y},\mathbf{z_2}) \approxeq \mathbf{I} - \mathbf{I}_{\mathbf{z_1}} \tag{5}$$

Since $\mathbf{I}_{\mathbf{y},\mathbf{z_2}}$, the reconstruction-error image, is not influenced by the nuisance factors related to $\mathbf{z_1}$, it can be used instead of $\mathbf{I}$ to train a PAD system.

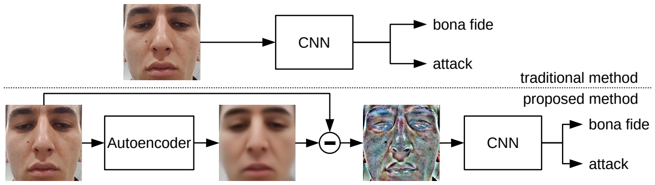

Fig. 2: Diagram of the proposed method. The upper part depicts the traditional approach of training a PAD system, where the original face images are used to train a CNN PAD system. In the proposed method, shown in the lower part of the figure, the autoencoder is first trained to reconstruct faces, using a large FR dataset. Then the reconstruction-error images computed from the output of this autoencoder are used to train the (CNN) PAD system. The autoencoder is not updated when the PAD-CNN is trained. The CNN is trained on reconstruction-error images of a PAD dataset.

Functions e and g can be modeled in an unsupervised manner using an *information maximizing variational autoencoder* (Info-VAE) [28]. The encoder and decoder parts of the autoencoder approximate e and g, respectively. Info-VAEs are able to learn *meaningful and disentangled* representations of samples where each dimension in the learned representation can represent one factor present in the data [28]. Info-VAEs learn these representations by imposing a prior distribution on their latent variables. By training an Info-VAE to reconstruct face images using only BF samples of an FR dataset, the autoencoder will model $\mathbf{z_1}$ as its latent variable. The trained autoencoder, when tested against BF and PA face images of a PAD dataset, will reconstruct face images only in terms of factors that it has modeled. In other words, the encoder, e, encodes any face image to its learned factors, $\mathbf{z_1}$, and the decoder reconstructs the face image using only those factors. The diagram of the proposed method is shown in Figure 2. The proposed method adds a pre-processing step using a pre-trained autoencoder to the PAD system compared to traditional methods. Instead of using original face images as input to a PAD system, we use the reconstruction error image of the autoencoder. Some examples of reconstruction error images are shown in Figure 3. We may observe that reconstruction error images look more similar to each other compared to the original images; The reconstruction error images are similar to each other in terms of color, contrast and so on. This is due to the removal of some of the nuisance factors.

4. IMPLEMENTATION DETAILS

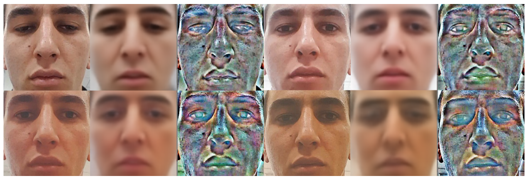

Fig. 3: Examples of autoencoder reconstruction-error images. The images in each three columns, from left to right, are original images, reconstructed images by the autoencoder, and reconstruction error images. The original image in top left is a BF sample and the rest of original images are PAs. The reconstruction error images contain less nuisance variations compared to the original images.

The following face PAD datasets that have been used in this study: OULU-NPU [7], Replay-Mobile [6], SWAN [29], and WMCA [8]. Only print and replay attacks are considered from each dataset. For the experiments discussed in Section 5, all models have been trained on OULU-NPU and evaluated on all four PAD datasets. The classification performance is reported in terms of area-under-the-curve (AUC) of *log-scale* receiver operating characteristic (ROC) curves. The ROC curves are computed with false positive rate ($APCER$ in [30]) along the x-axis (log-scale) and true positive rate ($1-BPCER$ in [30]) on the y-axis.[2] The proposed method is tested against the DeepPixBiS CNN architecture [9]. The architecture for the encoder part of the autoencoder is a DenseNet-161 [31] and the architecture of the decoder is a slightly modified version of the face generator in [32]. The size of $\mathbf{z_1}$, the latent variable of the autoencoder, is chosen to be 256 and its prior is arbitrarily assumed to be a Gaussian distribution with mean 0 and standard deviation of 3 (diagonal covariance matrix). The autoencoder is trained using cleaned versions (gray-scale images and images of statues were removed) of Microsoft Celeb (MS-Celeb-1M) [21] and the Celeb-A [33] FR datasets jointly.

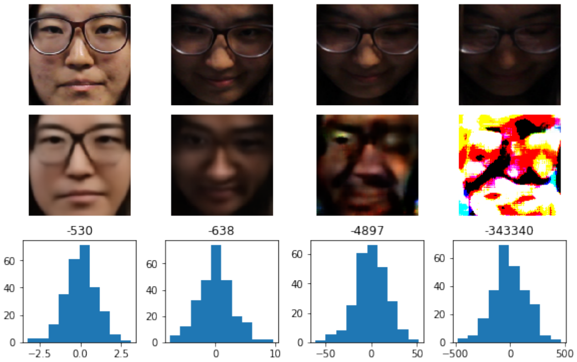

Fig. 4: Examples of the reconstructions of the autoencoder. Each row, from top to bottom, shows the original image, reconstruction by Info-VAE, and the histogram of the latent variables $\mathbf{z_1}$. Note that the ranges are different for the four histograms shown here. From left to right, the log-likelihood values of $\mathbf{z_1}$ given the prior distribution are: -530, -638, -1616, -4897, and -343340.

The reconstructed images generated by the autoencoder resemble low-pass filtered versions of the original images. Consequently, the reconstruction-error images will mainly contain the high frequency information of the original image. However, this approach is different from directly extracting high frequency components of the image based on a Gaussian-blur filter. To show the difference, we also compare our method with a PAD system that is trained on difference images between blurred images and original images.

[2] Note that since AUC of *log-scale* ROCs are reported, their values can be higher than 1.

This system will be called *Blur Error* in the experiments. In total, we will compare four methods:
1. *DeepPixBiS* [9]: our baseline PAD CNN.
2. *Blur Error* (BE): similar to the baseline but the input image $\mathbf{I}$ is first blurred using a Gaussian kernel, and the difference-image ($\mathbf{I}-\mathbf{I}_{\mathbf{blurred}}$) is used to train the baseline CNN.
3. *Autoencoder Error* (AE): like the baseline but the input images to DeepPixBiS are the reconstruction-error images of a pre-trained autoencoder.
4. *Thresholded Autoencoder Error* (TAE): similar to *AE* and is detailed below.

In the *TAE* method, input (test) images that do not meet certain quality criteria are rejected (not processed or scored) by the PAD system. In preliminary experiments we observed that the autoencoder cannot adequately reconstruct certain face images, such as very dark faces or faces with extremely non-frontal poses. Since the prior distribution for $\mathbf{z_1}$ is known, we can use the likelihood of each sample as a quality metric to reject unusual input images. If the likelihood of a sample is too small, the autoencoder is not able reconstruct the face image correctly since the decoder has not seen $\mathbf{z_1}$ values outside of the prior distribution. Figure 4 shows some face-image examples, the corresponding reconstructions and log-likelihood values. In our experiments we have set the threshold for the log-likelihood at -600 for rejecting samples. This threshold has been selected based on manual inspection of results in preliminary experiments. Overall, after thresholding the face images and rejecting some frames in videos, between 5% to 19% of videos were rejected, depending on the test dataset.

5. EXPERIMENTS

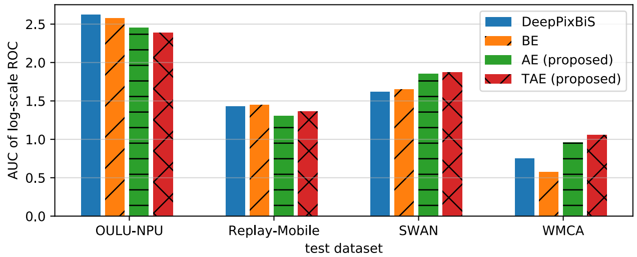

Fig. 5: Performance evaluation of the proposed method. The higher the value the better is the performance of the system. The dataset that the model was tested on is shown on the horizontal-axis.

Results of evaluating the various networks on different datasets are shown in Figure 5. The various datasets are shown on the horizontal-axis. As the OULU-NPU dataset has been used to train the networks, results for this dataset correspond to the intra-dataset evaluations. The best performing method is the *DeepPixBiS* baseline in intra-dataset evaluations. The performance of the BE method is slightly lower than that of *DeepPixBiS*, and the performance of the proposed AE and TAE methods are even worse. However, we argue that the baseline methods are overfitting on the OULU-NPU dataset in intra-dataset evaluations as their performance degrades significantly in the cross-dataset scenarios, that is, when the test dataset is not OULU-NPU. Overall, we can see that the TAE method performs slightly better than the AE method in all cross-dataset evaluations. For the SWAN and the WMCA datasets, both proposed methods (AE and TAE) perform significantly better than the baselines. For the Replay-Mobile dataset, however, all methods show similar performance, and the proposed AE method performs slightly worse compared to the baselines.

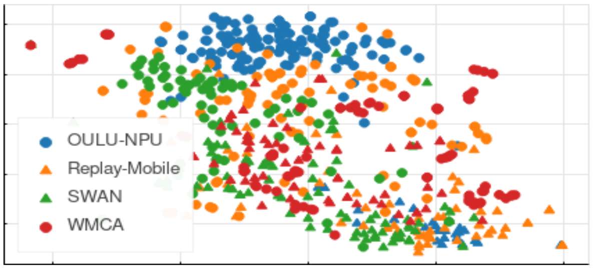

Fig. 6: t-SNE plot of the embeddings of the *AE* system similar to Figure 1.

We have investigated the low performance of proposed AE method in the case of Replay-Mobile. In this dataset, some face images are either very dark or have very strong lateral illumination. These samples are annotated with lighting condition of *adverse* and *lateral* in the dataset. Most of the classification errors of the AE method correspond to such samples. As discussed before, we found this problem to be mainly due to bad reconstructions of the autoencoder. In fact the TAE method, in which bad reconstructions were removed, had similar performance compared to the baselines when evaluated on Replay-Mobile.

Figure 6 shows a 2D t-SNE plot for embeddings produced by the AE method in a fashion similar to that used in Figure 1. We observe in the plot that, unlike in Figure 1, embeddings corresponding to samples of the BF class (triangles) from the different datasets are mixed together. For the PA class (circles), however, the embeddings still form fairly compact clusters by dataset. One reason for this phenomenon may be the following. In this work, we have not explicitly tried to suppress the effect of $\mathbf{z_2}$, the latent variable representing the ensemble of nuisance factors exclusive to PAs. Therefore, the factors influencing $\mathbf{z_2}$ may still cause the PA embeddings to form compact clusters.

6. CONCLUSIONS

We have presented a novel approach to improving generalization of face PAD by taking advantage of large public FR datasets which contain millions of BF face images from varied sources. We hypothesize that all the factors (variability) present in face images of an FR dataset are nuisance factors to PAD systems. By explicitly modeling these factors using an Info-VAE (an autoencoder which learns meaningful and disentangled representations of data), we induce invariance of these factors to a PAD system. This is done by reconstructing face images with the pre-trained autoencoder and using the reconstruction-error image (the difference between the original image and the reconstructed one) as input to the face-PAD system. We assume here that the face image reconstructed by the autoencoder only contains information about the nuisance factors of PAD. When the baseline PAD system is trained on the reconstruction-error images, the intra-dataset performance degrades slightly, but the cross-dataset performance improves significantly for two out of three test datasets. This supports our hypothesis that the influence of some nuisance factors on a face-PAD system can be lowered by incorporating knowledge from an FR dataset. Furthermore, using the Info-VAE allows us to systematically reject low quality samples, which also contributes to the improved performance.

7. REFERENCES

[1] David Sandberg, "Facenet: Face recognition using tensorflow," 2017.

[2] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, vol. 1, p. 6.

[3] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light CNN for deep face representation with noisy labels," *arXiv preprint arXiv:1511.02683*, 2015.

[4] Amir Mohammadi, Sushil Bhattacharjee, and Sébastien Marcel, "Deeply vulnerable: A study of the robustness of face recognition to presentation attacks," *IET Biometrics*, vol. 7, no. 1, pp. 15–26, 2017.

[5] "Information technology – Biometric presentation attack detection – Part 1: Framework," Standard, International Organization for Standardization, Geneva, CH, Jan. 2016.

[6] Artur Costa-Pazo, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sebastien Marcel, "The REPLAY-MOBILE Face Presentation-Attack Database," in *Biometrics Special Interest Group (BIOSIG), 2016 International Conference of The*. 2016, pp. 1–7, IEEE.

[7] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference On*. 2017, pp. 612–618, IEEE.

[8] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel, "Biometric Face Presentation Attack Detection with Multi-Channel Convolutional Neural Network," *IEEE Transactions on Information Forensics and Security*, 2019.

[9] Anjith George and Sébastien Marcel, "Deep Pixel-wise Binary Supervision for Face Presentation Attack Detection," in *International Conference on Biometrics*, 2019.

[10] Xiaoguang Tu, Jian Zhao, Mei Xie, Guodong Du, Hengsheng Zhang, Jianshu Li, Zheng Ma, and Jiashi Feng, "Learning Generalizable and Identity-Discriminative Representations for Face Anti-Spoofing," *arXiv preprint arXiv:1901.05602*, 2019.

[11] Xiaoguang Tu, Hengsheng Zhang, Mei Xie, Yao Luo, Yuefei Zhang, and Zheng Ma, "Deep Transfer Across Domains for Face Anti-spoofing," *arXiv preprint arXiv:1901.05633*, 2019.

[12] A. Gretton, AJ. Smola, J. Huang, M. Schmittfull, KM. Borgwardt, and B. Schölkopf, "Covariate shift and local learning by distribution matching," in *Dataset Shift in Machine Learning*, pp. 131–160. Biologische Kybernetik, Cambridge, MA, USA, 2009.

[13] Mei Wang and Weihong Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[14] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig, "Controllable invariance through adversarial feature learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 585–596.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[17] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[18] Yujia Li, Kevin Swersky, and Richard Zemel, "Learning unbiased features," *arXiv preprint arXiv:1412.5244*, 2014.

[19] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan, "Unsupervised adversarial invariance," in *Advances in Neural Information Processing Systems*, 2018, pp. 5092–5102.

[20] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," *arXiv preprint arXiv:1607.08221*, 2016.

[22] Xiaowen Ying, Xin Li, and Mooi Choo Chuah, "LiveFace: A Multi-task CNN for Fast Face-Authentication," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018, pp. 955–960, IEEE.

[23] Ayush Jaiswal, Shuai Xia, Iacopo Masi, and Wael AbdAlmageed, "RoPAD: Robust Presentation Attack Detection through Unsupervised Adversarial Invariance," *arXiv preprint arXiv:1903.03691*, 2019.

[24] Robbie Vogt and Sridha Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[25] Roy Wallace, Mitchell McLaren, Christopher McCool, and Sebastien Marcel, "Inter-session variability modelling and joint factor analysis for face authentication," in *Biometrics (IJCB), 2011 International Joint Conference On*. 2011, pp. 1–8, IEEE.

[26] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[27] Fabien Cardinaux, Conrad Sanderson, and Samy Bengio, "User authentication via adapted statistical models of face images," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 361–373, 2006, 00118.

[28] Shengjia Zhao, Jiaming Song, and Stefano Ermon, "InfoVAE: Information Maximizing Variational Autoencoders," *arXiv:1706.02262 [cs, stat]*, June 2017.

[29] Raghavendra Ramachandra, Martin Stokkenes, Amir Mohammadi, Sushma Venkatesh, Kiran Raja, Pankaj Wasnik, Eric Poiret, Sébastien Marcel, and Christoph Busch, "Smartphone Multi-modal Biometric Authentication: Database and Evaluation," *arXiv:1912.02487 [cs]*, Dec. 2019.

[30] "ISO/IEC DIS 30107-3. Information Technology – Biometric presentation attack detection – Part 3: Testing and reporting," Standard, International Organization for Standardization, Geneva, CH, Jan. 2016.

[31] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[32] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, Dec. 2015.