EG UK Computer Graphics & Visual Computing (2021)
K. Xu and M. Turner (Editors)

Training dataset construction for anomaly detection in face anti-spoofing

L. Abduh and I. Ivrissimtzis

Durham University, Department of Computer Science, UK

Abstract

Anomaly detection, which is approaching the problem of face anti-spoofing as a one-class classification problem, is emerging as an increasingly popular alternative to the traditional approach of training binary classifiers on specialized anti-spoofing databases which contain both client and imposter samples. In this paper, we discuss the training protocols in the existing work on anomaly detection for face anti-spoofing, and note that they use images exclusively from specialized anti-spoofing databases, even though only common images of real faces are needed.

In a proof-of-concept experiment, we demonstrate the potential benefits of adding in the anomaly detection training sets images from general face recognition, rather than specialised face anti-spoofing, databases, or images from the in-the-wild images. We train a convolutional autoencoder on real faces and compare the reconstruction error against a threshold to classify a face image as either client or imposter. Our results show that the inclusion in the training set of in-the-wild images increases the discriminating power of the classifier on an unseen database, as evidenced by an increase in the value of the Area Under the Curve.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision tasks;*** *Image manipulation;*

1. Introduction

Face liveness tests authenticate users of face recognition systems by processing input images and deciding whether they come from a human face or, for example, from printed photos held in front of the system's camera by an imposter. The main challenge for developing a robust face anti-spoofing system is the large number of different types of *presentation attacks* the system must learn to recognize. For example, an imposter could be presenting to the face recognition system a printed photo, a screen displaying a still image, or a screen replaying a video. A multitude of other factors, such as the quality of the printed photo, the resolution and type of the displaying screen, the illumination conditions of the scene, and the characteristics of the system's camera, may also have a significant effect on the performance of any anti-spoofing algorithm. Moreover, a robust anti-spoofing algorithm should be able to cope with previously unseen attack methods, which were not anticipated prior to its deployment.

Traditionally, face anti-spoofing is approached as a binary classification problem and classifiers are trained on specialised datasets, containing both client and imposter images and videos. The main limitation of this approach is associated with the high cost of creating such databases. That is, a limited only number of attacks is simulated, on a limited number of subjects, while the variability of important environmental factors such as illumination conditions and background is also limited. As a result, the classifiers do not always generalize well to previously unseen attacks.

In this context, anomaly detection, using classifiers trained on a one class dataset of client images only, is becoming an increasingly popular approach to face anti-spoofing [AKC17] [AK18]. The present work is motivated by the observation that training with client images only can also use in-the-wild face images, that is, a set of face images harvested online, as well as face images from databases that do not specialize in face-anti-spoofing.

After giving a brief overview of the general literature on face anti-spoofing, in Section 2.2 we review the relevant literature on the use of anomaly detection for face anti-spoofing and establish our main observation. That is, in the existing literature, the training data are drawn from specialised face anti-spoofing databases, even though they are just common face images.

In Sections 3 and 4, we describe a proof-of-concept experiment on the feasibility of an alternative approach to the creation of one-class training sets. In particular, we augment an initial training set of client images from specialised face anti-spoofing databases, first with images from non-specialised databases, the SCFace [GDG11] and the CASIA-Web Face [YLLL14] in particular, and then with images from the in-the-wild, which were semi-automatically harvested from online sources.

DOI: 10.2312/cgvc.20211312

Our anomaly detection anti-spoofing algorithm is based on a Convolutional Autoencoder (ACE), similar to the one we used in [AI20]. Following a well-established methodology, the ACE is trained on client images, and test images are classified as clients when their reconstruction error is below a threshold. First, we trained the ACE with client images from the Replay-Attack [CAM12] database, and tested it on the Replay-Attack and NUAA [TLLJ10] databases, creating a baseline. Next, we added into the training dataset the images from the in-the-wild, which were semi-automatically collected from online sources, and finally, we added to the training set images from SCFace and the CASIA-Web Face, which do not specialize in face anti-spoofing. The results show that the classifier's discriminative power, as measured by the Area Under the Curve metric, increases markedly on the unseen NUAA, with a moderate only drop on Replay-Attack. Finally, we added to the training set images from databases that do not specialize in anti-spoofing, SCFace [GDG11] and CASIA-Web face [YLLL14] in particular, obtaining again similar results.

The main contributions of the paper are:

- We review the literature on anomaly detection for face anti-spoofing and establish the observation that the training sets consist of images drawn from specialised face anti-spoofing databases.
- In a proof-of-concept experiment, we developed an anomaly detection method for face anti-spoofing based on a convolutional autoencoder and tested it on the previously unseen NUAA database, showing performance increases when we add into the training set in-the-wild face images and face images from non-specialized databases.

2. Background

The progress in the field of face anti-spoofing is inextricably linked to the development of specialised, image and video, anti-spoofing databases. The first such database to become publicly available, and set the standards for subsequent developments, was the NUAA Photograph Imposter Dataset [TLLJ10]. The NUAA samples were collected from 15 subjects, using a cheap webcam, in three sessions on different environments and illumination conditions. The attacks were based on digital images captured with a professional camera, and then printed on paper at various resolutions. Other notable databases that are most commonly used in the literature, are the Replay-Attack [CAM12], which we will use here to form the baseline of our experiment, the CASIA-FASD [ZYL*12], and the MSU-MFSD [WHJ15]. These newer databases include some novel characteristics, especially the video presentation attacks, however, regarding the number of subjects, the number of different types of simulated attacks and the variability of the environmental conditions, they do not differ materially from the NUAA.

2.1. Face anti-spoofing

In the past few years, a large number of methods have been proposed for the face presentation attack problem (PAD). Such methods can be classified into intrusive and non-intrusive types [PAHA15], depending on their interference with the bio-metric data acquisition process. The non-intrusive methods have received more attention in the literature. Another categorization of spoofing detection methods examines the way the classification algorithm handles image features. On the one hand, we have the traditional face anti-spoofing methods, which use hand-crafted features and employ shallow machine learning, and on the other hand the deep learning methods.

Regarding the more traditional approaches to anti-spoofing, [TLLJ10] studied several hand-crafted feature / shallow classifier combinations. The features included Differences of Gaussians, and features obtained through Logarithmic Total Variation smoothing, while their classifiers included Sparse Logistic Regression, Sparse Low Rank Bilinear Logistic Regression, and SVMs. In subsequent work, Local Binary Patterns (LBPs) are the most commonly used image features. In [CAM12], LBPs are used against various presentation attacks, such as printed photographs, digital photos and videos.

The above shallow methods do not always generalise well to previously unseen attacks. Deep learning is an alternative approach, which regularly outperforms more traditional approaches, since, in the context of such complex tasks, multi-layered methods seem better suited for the extraction of the high-level features of a dataset [WHJ15].

Convolutional Neural Networks (CNNs) in particular have achieved impressive results on a range of image and video classification tasks. One of the earliest attempts on face anti-spoofing with CNNs is Yang et al. [YLL14]. Their results were improved by Atoum et al. [ALJL17] using a two-stream CNN-based network, which performed well under an intra-dataset testing protocol. Xu et al. [XLD15] extract temporal elements using a deep architecture combining LSTM units with convolutional layers and max-pooling. Again, their model performed well under an intra-dataset testing protocol, however, as it was usually the case with the early deep learning approaches, cross-database generalisability was poor. The early algorithms did not perform well on previously unseen attacks, and also could not cope with the high cross-database variability of the environmental conditions, as well as variability in image and video quality.

Going beyond the concise literature review on PAD approaches included in this paper for self-containment and completeness, for a more detailed review of the area we refer the reader to the recent survey in [MVLB20]. While it covers only PAD methods that utilize common RGB cameras on consumer level devices, it contains a comprehensive overview of the relevant publicly available databases, as well as substantial experimental results comparing the various PAD methods.

2.2. Anomaly detection in face anti-spoofing

In [XA18], the proposed anomaly detection classifier uses an autoencoder for feature extraction, followed by a one-class SVM for classification. The validation of the method focuses on its performance on previously unseen attacks, rather than previously unseen databases. In the various experiments, the network was trained using client samples from the CASIA, Replay-Attack and MSU databases. To demonstrate results were presented for networks trained on subsets of the initial training set, but in all cases only

client images from specialised presentation attack databases were used.

In [YLM16], they use LBPs to extract low-level features, followed by a sparse autoencoder extracting high-level features, while the final classification was based on a LibSVM. The training and the testing utilised the CASIA and the NUAA databases.

In [AKC17], they use again hand-crafted features, such as LBPs and image quality metrics, and then they show that one-class classifiers work better than the binary ones on cross-database testing mode. Training and testing utilises three specialised databases, the CASIA-FASD, Replay-Attack, and the MSU-MFSD.

In [AK18], they use one-class classifiers and demonstrate that the utilisation of subject-specific information can improve considerably the system's performance. Training and testing utilised an aggregated composition of three specialised databases, Replay-Attack, Replay-Mobile, and MSU-MFSD.

In [NMAM18], assuming that client images have similar texture types, a Gaussian Mixture Model is used to learn textures. Again, one-class classifiers were shown to outperform the binary ones. Training and testing utilised the Replay-Attack database and the enrolment data of each client.

In [JCPC19], the anomaly detection classifier is based on the use of a deep metric learning model with a triplet focal loss as regulariser. They trained and tested on subsets of the GRAD-GPAD, which is an aggregated dataset from several specialised anti-spoofing datasets, and their protocols took explicitly in consideration the type of the attacks, lightning conditions, capturing device and resolution.

In [BOPP20], they used one-class CNNs trained by client images only. The authors noted the close proximity between the client and imposter feature spaces, and they introduced adaptive mean estimation strategy to generate pseudo-negative data following a Gaussian distribution. They trained and tested on four specialised datasets: Replay-Attack, Rose-Youtu, OULU-NPU and Spoof-in-Wild.

The approach in [NGM19] differs from the ones discussed above in that depth and near-infrared information is utilised together with RGB images from the WMCA dataset. We note, however, the use of the non-specialised CelebA database to train a convolutional autoencoder, before an MLP is trained as a binary classifier, with both client and imposter images.

3. Experiment

Our proof-of-concept experiment is based on a convolutional autoencoder. Autoencoders are neural networks consisting of two parts. The *encoder* processes the input image and produces the *code*, a compressed representation of the input which, usually, has a much lower dimension. The *decoder* reconstructs the original image from the code. The loss function is the reconstruction error, here the Mean Squared Error (MSE) between the original Y and the reconstructed image $\hat{Y}$, seen as $3mn$-dimensional vectors, i.e., one dimension per pixel, per colour band.

$$\mathrm{MSE} = \frac{1}{3mn}\sum_{k=1}^{3}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(Y_{i,j,k} - \hat{Y}_{i,j,k}\right)^2 \qquad (1)$$

As the network is trained to minimise the reconstruction error of the client images, a high reconstruction error indicates images outside the client class. Thus, we classify an image as imposter when the reconstruction error is higher than a predefined threshold.

Following [XCW*15], we implemented both the encoder and the decoder as multi-layer CNNs. Following [BRF18], the entire autoencoder was trained with images from a single class, here the client images.

3.1. The autoencoder

Figure 1 shows the architecture of the proposed autoencoder. The input is a 64×64 RGB image; the encoder consists of three convolution layers of kernel size (3,3), each one followed by a Max-Pooling layer of kernel size (2,2), which is used for spatial downsampling. The decoder consists of two transpose convolutional layers, followed by one convolution layer, and reconstructs a representation of original input image. In all layers, we used ReLu activation functions, except for the last layer where a sigmoid function was used.

All code was written in Python, on the Keras platform, and the experiments ran on an Intel Core i7 CPU 64 GB RAM PC with an Nvidia GTX 1650. The whole network was trained with the RMSprop optimizer for 50 epochs, with a learning rate of 0.001. The batch size was set to 32.

3.2. Training, validation, and test datasets

The faces in all training and testing images were detected with the Haar feature-based cascade classifier [VJ01], followed by manual inspection and selection. The user input was necessary, especially in the creation of training images from the in-the-wild, due to performance issues of the face detector on such images; general image quality issues such as out of focus blurry faces; and in few cases by the need to exclude imposter images, e.g. faces on a poster on a wall. All selected face images were cropped and normalized to 64×64 pixels. We note that face detection followed by cropping is a standard procedure in PAD. In fact, as a standard practice, the images in the benchmark databases are accompanied by a set of coordinates giving the positions of the faces.

We tested the autoencoder on two test datasets, the first from the Replay-Attack and the other from the NUAA, consisting of 236 images each. The imposter subset contained images from all types of attacks supported by these two databases. We note that NUAA is consider a particularly challenging case when one is testing for cross-database generalization [YLM16], and the use of webcams as capture devices increases further the challenge of generalisability.

The aim of this paper is not to propose an optimised network architecture, but we focus instead on studying the effect of the training set on the generalisability. Thus, the architecture and the training protocol of the autoencoder are fixed, and the main variable of our experiment is the training set. To see how the augmentation of the training set with images from non-specialised databases affects the generalisation power of the classifier across the two test sets, we opted for three training sets such that D1 is a subset of D2 and D3, and D2 subset of D3:

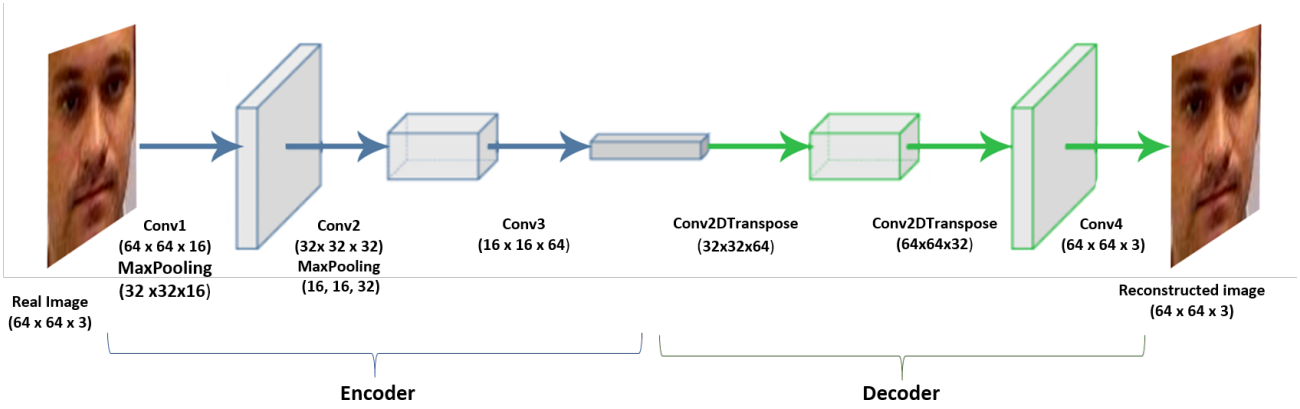

Figure 1: *The architecture of the proposed convolutional autoencoder.*

Table 1: *Description and size of the training datasets.*

	Description	size
D1	Replay-Attack	2800
D2	Replay-Attack + in-the-wild mages	2902
D3	Combine Replay-Attack with others DB	3422

D1. Images form the Replay-Attack dataset only. We used 10 client subjects' videos, both controlled and adverse.
D2. We added to D1 102 face images harvested online using general keywords such as *teachers*. These 163 face images were manually chosen from a larger collection, the main considerations being to be frontal face images, in-focus, and of a good size so the normalisation to size 64×64 does not require excessive zooming.
D3. We added 520 images from the SCFace [GDG11] and the CASIA-WebFace databases. The SCFace is a surveillance camera face database from which we used the mugshot, still color images, captured indoors under controlled illumination conditions. The CASIA-WebFace is a very large dataset, consisting of 10,575 subjects, collected in a semi-automatic way from the Internet. We used a random subset of it.

Table 1 summarizes the description of the training datasets.

The validation dataset was kept constant to simplify the design of the experiment. It consisted of 578 live and fake images from the Replay-Attack. As the use of a validation set with a composition similar to the most general training dataset **D3** may lead to an underestimation of the performance of proposed autoencoder on the Replay-Attack under an intra-database protocol, we also report HTER values computed under the use of a validation set consisting of Replay-Attack images only.

4. Results and discussion

Figure 2 shows the ROC curves of the proposed autoencoder, trained on the three datasets, and tested on Replay-Attack (left) and NUAA (right). The corresponding Areas Under the Curve(AUC) are reported on Table 2. We notice that the inclusion of the in-the-wild images in the training dataset improved markedly the cross-database generalisation power of the classifier, with the value of the AUC on the NUAA going up from 0.63 when trained with **D1** to 0.72 when trained with **D2**. Moreover, the inclusion of images from non-specialized databases further increased further the AUC to 0.80 when trained with **D3**. There is also a noticeable fall on the performance on Replay-Attack, with the AUC going down from .93 to .89 and then to .82. We also note the high performance of the algorithm under an intra-database test mode, that is, the high AUC value of .93 AUC on the Replay-Attack.

Table 2: *The AUC values corresponding to the ROC curves shown in Figure 2.*

	D1	D2	D3
Replay-Attack	.93	.89	.82
NUAA	.63	.72	.80

The value of the AUC is an integral over all possible operating points, that is, overall possible thresholds against which we compare the reconstruction error to determine whether a sample should be classified as client or imposter. Thus, it separates the problem of assessing the discriminative power of the classifier from the problem of finding an optimal, for the given test, operating point. Next, we will discuss the problem determining an optimal operating point.

In the literature, classifier performance on a specific operating point is usually assessed either by reporting separately the False Positive Rate (FPR) and the False Negative Rate (FNR), or their mean average Half Total Error Rate. We note that reporting an operating specific performance metric does not necessarily mean that the problem of finding the optimal operating point has been addressed. For example, some papers report the minimum HTER over all operating points, or the True Positive Rate corresponding to certain fixed values of FPR. Employing a technique that is commonly used to address this problem, we first compute a threshold on the validation set, here the threshold corresponding to the Equal Er-

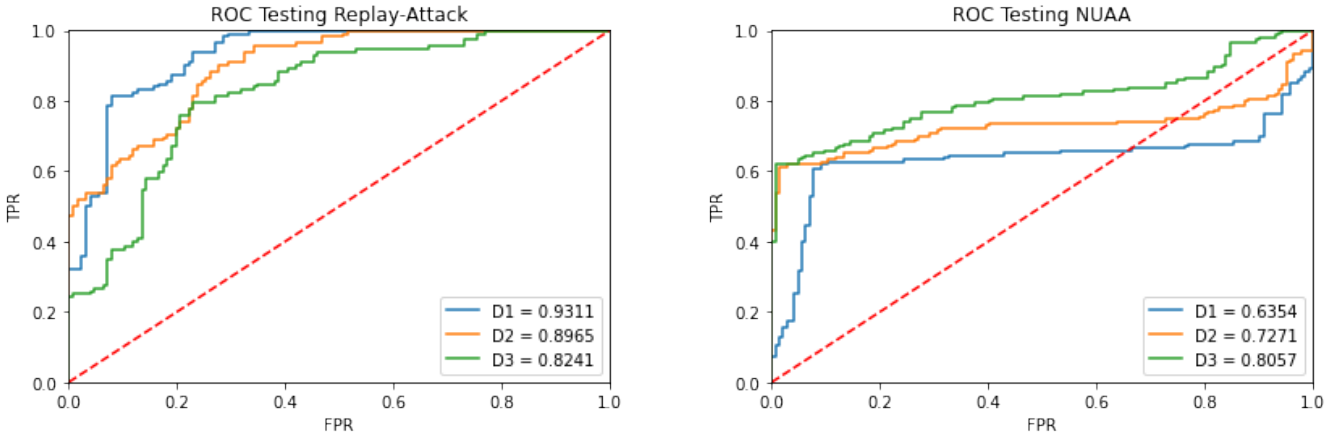

Figure 2: *ROC curves corresponding to classifier/training dataset combinations, tested on Replay-Attack (left) and NUAA (right).*

ror Rate (EER) on that set, and use this threshold to compute the HTER.

Table 3 summarizes the HTERs of our method, and for comparison, the HTER range of the four different classifiers proposed in [BdSPR*19]. Regarding the performance on the NUAA, despite the satisfactory discriminating power of the classifier as shown by the ROC curves, the high HTER values indicate that the threshold computed on the validation set, which is a set containing images from the Replay-Attack only, cannot be used on the NUAA. We note that [BdSPR*19] also reports very high HTERs, which again indicate that a satisfactory operating point on NUAA could not be found. Regarding the performance on the Replay-Attack, we note that under such an intra-database testing mode, our convolutional autoencoder performs significantly worse on Replay-Attack than the best performing classifier in [BdSPR*19], but its performance is inside their reported range.

Table 3: *HTERs computed on the operating point corresponding to the EER of the validation set. The last column shows the range of HTERs reported in [BdSPR*19].*

	D1	**D2**	**D3**	[BdSPR*19]
Replay-Attack	.15	.24	.26	[.05 - .32]
NUAA	.57	.54	.39	[.51 - .65]

5. Conclusions and future work

A research of the literature on anomaly detection methods for face anti-spoofing shows that their training datasets are exclusively drawn from specialised anti-spoofing databases. This seems to be an unnecessary limitation, given that the one-class training sets of the anomaly detection methods requires just normal face images, which can be found in abundance in non-specialised databases, or can be obtained from the in-the-wild.

In a proof-of-concept experiment, we showed that the inclusion of such images in the training set of a convolutional autoencoder, which was originally trained on the Replay-Attack database, increased its performance on the unseen NUAA database, as shown by the ROC curves and the corresponding AUC values while, conversely, the performance on the Replay-Attack itself decreased. That is, as expected, in a cross-database testing mode, the inclusion of images from outside the specialised databases had a moderating effect, rather than a positive or a negative one.

We note that in the most recent papers on face anti-spoofing, as the ones reviewed here, cross-database testing is becoming the norm. That is, regardless of its relevance in typical practical application scenarios, in the literature, the issue of how the classifier would perform on images from previously unseen databases is considered important. In this context, the behaviour of the classifier on unseen client images outside specialised anti-spoofing databases becomes an equally legitimate question. The various methodological challenges that would arise from the use of such non-symmetric test sets, where the client images will be drawn from more sources than the imposter images, is an issue we plan to address in our future work.

To increase further the scope of our investigation, in the future, we also plan to work on evaluating the performance of the various training sets in conjunction with anomaly detection classifiers based on adversarial models, such as BiGANs [ZFL*18], AnoGANs [SSW*17], which, recently, have been employed to tackle the PAD problem [GNO19].

References

[AI20] ABDUH L., IVRISSIMTZIS I.: Use of in-the-wild images for anomaly detection in face anti-spoofing. *arXiv 2006.10626* (2020). 2

[AK18] ARASHLOO S. R., KITTLER J.: Client-specific anomaly detection for face presentation attack detection. *arXiv 1807.00848* (2018). 1, 3

[AKC17] ARASHLOO S. R., KITTLER J., CHRISTMAS W.: An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access 5* (2017), 13868–13882. 1, 3

[ALJL17] ATOUM Y., LIU Y., JOURABLOO A., LIU X.: Face anti-spoofing using patch and depth-based cnns. In *Proc. IJCB* (2017), IEEE), pp. 319–328. 2

[BdSPR*19] BRESAN R., DA SILVA PINTO A., ROCHA A., BELUZO C., CARVALHO T.: Facespoof buster: a presentation attack detector based on intrinsic image properties and deep learning. *arXiv 1902.02845* (2019). 5

[BOPP20] BAWEJA Y., OZA P., PERERA P., PATEL V. M.: Anomaly detection-based unknown face presentation attack detection. In *2020 IEEE International Joint Conference on Biometrics (IJCB)* (2020), IEEE, pp. 1–9. 3

[BRF18] BHATTAD A., ROCK J., FORSYTH D.: Detecting anomalous faces with 'no peeking' autoencoders. *arXiv 1802.05798* (2018). 3

[CAM12] CHINGOVSKA I., ANJOS A., MARCEL S.: On the effectiveness of local binary patterns in face anti-spoofing. In *Proc. BIOSIG* (2012), IEEE), pp. 1–7. 2

[GDG11] GRGIC M., DELAC K., GRGIC S.: Scface–surveillance cameras face database. *Multimedia tools and applications 51*, 3 (2011), 863–879. 1, 2, 4

[GNO19] GUPTA V., NISHIGAKI M., OHKI T.: Unsupervised biometric anti-spoofing using generative adversarial networks. *International Journal of Informatics Society 11*, 1 (2019). 5

[JCPC19] JIMÉNEZ-CABELLO D., PÉREZ-CABO D.: Deep anomaly detection for generalized face anti-spoofing. In *Actas del IV Machine Learning Workshop* (2019), University of A Coruña, pp. 1–31. 3

[MVLB20] MING Z., VISANI M., LUQMAN M. M., BURIE J.-C.: A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices. *Journal of Imaging 6*, 12 (2020). URL: https://www.mdpi.com/2313-433X/6/12/139, doi:10.3390/jimaging6120139. 2

[NGM19] NIKISINS O., GEORGE A., MARCEL S.: Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In *2019 International Conference on Biometrics (ICB)* (2019), IEEE, pp. 1–8. 3

[NMAM18] NIKISINS O., MOHAMMADI A., ANJOS A., MARCEL S.: On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In *Proc. ICB* (2018), IEEE), pp. 75–81. 3

[PAHA15] PARVEEN S., AHMAD S. M. S., HANAFI M., ADNAN W. A. W.: Face anti-spoofing methods. *Current science* (2015), 1491–1500. 2

[SSW*17] SCHLEGL T., SEEBÖCK P., WALDSTEIN S. M., SCHMIDT-ERFURTH U., LANGS G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging* (2017), Springer, pp. 146–157. 5

[TLLJ10] TAN X., LI Y., LIU J., JIANG L.: Face liveness detection from a single image with sparse low rank bi-linear discriminative model. In *Proc. ECCV* (2010), Springer), pp. 504–517. 2

[VJ01] VIOLA P., JONES M.: Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR* (2001), vol. 1, IEEE), pp. I–I. 3

[WHJ15] WEN D., HAN H., JAIN A. K.: Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security 10*, 4 (2015), 746–761. 2

[XA18] XIONG F., ABDALMAGEED W.: Unknown presentation attack detection with face rgb images. In *Proc. BTAS* (2018), IEEE), pp. 1–9. 2

[XCW*15] XIA Y., CAO X., WEN F., HUA G., SUN J.: Learning discriminative reconstructions for unsupervised outlier removal. In *Proc. ICCV* (2015), IEEE), pp. 1511–1519. 3

[XLD15] XU Z., LI S., DENG W.: Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *Proc. ACPR* (2015), IEEE), pp. 141–145. 2

[YLL14] YANG J., LEI Z., LI S. Z.: Learn convolutional neural network for face anti-spoofing. *arXiv 1408.5601* (2014). 2

[YLLL14] YI D., LEI Z., LIAO S., LI S. Z.: Learning face representation from scratch. *arXiv 1411.7923* (2014). 1, 2

[YLM16] YANG D., LAI J., MEI L.: Deep representations based on sparse auto-encoder networks for face spoofing detection. In *Proc. Chinese Conference on Biometric Recognition* (2016), Springer), pp. 620–627. 3

[ZFL*18] ZENATI H., FOO C.-S., LECOUAT B., MANEK G., CHANDRASEKHAR V.: Efficient gan-based anomaly detection. *ArXiv abs/1802.06222* (2018). 5

[ZYL*12] ZHANG Z., YAN J., LIU S., LEI Z., YI D., LI S. Z.: A face antispoofing database with diverse attacks. In *Proc. ICB* (2012), IEEE), pp. 26–31. 2