ISSN 1213-6972 (print)
ISSN 1213-6964 (online)
Journal of WSCG
http://www.wscg.eu
Vol.31, No-1-2, 2023

Anomaly Detection with Transformers in Face Anti-spoofing

Latifah Abduh
Department of Computer Science
Durham University
Durham, DH1 3LE, UK
0000-0002-6359-311X
latifah.a.abduh@dur.ac.uk

Luma Omar
Department of Computer Science
Durham University
Durham, DH1 3LE, UK
0000-0002-7215-9112
lama.omar@yahoo.com

Ioannis Ivrissimtzis
Department of Computer Science
Durham University
Durham, DH1 3LE, UK
0000-0002-3380-1889
ioannis.ivrissimtzis@dur.ac.uk

ABSTRACT

Transformers are emerging as the new gold standard in various computer vision applications, and have already been used in face anti-spoofing demonstrating competitive performance. In this paper, we propose a network with the ViT transformer and ResNet as the backbone for anomaly detection in face anti-spoofing, and compare the performance of various one-class classifiers at the end of the pipeline, such as one-class SVM, Isolation Forest, and decoders. Test results on the RA and SiW databases show the proposed approach to be competitive as an anomaly detection method for face anti-spoofing.

Keywords

Face presentation attack, Vision Transformer, ResNet, anomaly detection, one-class classification.

1 INTRODUCTION

While face recognition is the biometric authentication method of choice in many application domains, it is still considered extremely vulnerable to presentation attacks. In such attacks, an imposter is trying to gain unlawful access by presenting in front of the system's camera a printed photo, or an electronic screen playing a video of a rightfully registered person. The vulnerability of face recognition systems to such spoofing attacks means that they cannot be safely deployed in security-sensitive applications in uncontrolled environments, as for example ATM machines in the high street. Presentation attack detection (PAD) addresses this problem by developing binary classification algorithms aiming at distinguishing between the genuine, bona fide samples presented to the system's camera, and the imposter ones.

The most common approach to PAD is to train a binary classifier on both the bona fide and the imposter classes. In this case, training and testing are performed within specialised face anti-spoofing databases, which due to the high cost of producing imposter samples have limited variability, raising questions on the generalisation power of the classifier, especially on unseen attacks in scenarios that have not been covered by the testing databases. In particular, while the current state-of-the-art algorithms can show good results on unseen attacks within the same database, and some generalisation power between specific databases, a thorough cross-database validation is expected to show that they do not always generalise well. For example, in [1] all the eleven methods under comparison show HTERs between 24% and 60.6% in cross-database generalisation task from the Replay Attack database to the CASIA-MFSD.

An alternative approach aiming at addressing the generalisation problem is anomaly detection based on one-class training. We note that in the limited testing environments provided by the existing databases, anomaly detection approaches underperform two-class training under most testing protocols. However, they have the conceptually appealing property that they neither attempt to learn specific presentation attacks nor, most importantly, specific environments where such attacks where modelled during the creation of the database. Thus, anomaly detection for face anti-spoofing is still a very active research area [2, 3].

In this paper, we use the Vision Transformer (ViT) [4] and the ResNet [5] as backbones for anomaly detection for face anti-spoofing. Our motivation for using ViT was the observation that while in several computer vision tasks Transformers are replacing Convolutional Neural Networks (CNNs) as the new gold standard, and they have already been proposed for the PAD problem under a two-class training setting [6], they have not been used yet for PAD in the anomaly detection setting.

Regarding the use of ResNet, we note that the size of the receptive field is one of the primary distinctions between a CNN-based model and a transformer-based model. Whereas due to the self-attention mechanism, the transformer is superior in its ability to capture a pixel relation over a long distance [7], nonetheless, it

lacks a reliable way of capturing spatial information within each patch, so it may overlook a crucial spatial local patterns, such as textures. However, CNNs are different in this regard, focusing on textures rather than shapes to identify objects in images [8]. ResNet in particular is a highly efficient neural network architecture and its residual learning methodology addresses the degradation issue which exists in many other CNN models. Thus, overall, we leverage the strengths of two state-of-the-art architectures, a transformer and a CNN to extract reliable features. Our ablation study shows that the combined ViT ResNet backbone gives significant improvement over a single network backbone.

Our main contributions are summarised as follows:

- A novel Anomaly detection Vision Transformer (AnoFormer), with ViT and ResNet in the backbone, for presentation attack detection.
- A comparison of various one-class classification models, showing that a decoder with MSE as a loss function outperforms the other configurations.
- An ablation study showing that the use of a combination of ViT and ResNet in the backbone outperforms the use of single networks.

The rest of the paper is organised as follows. In Section 2 we review the related work. In Section 3 we present the proposed AnoFormer and its implementation details. In Section 4 we present the results, and we briefly conclude in Section 5.

2 RELATED WORK

2.1 Face Anti-spoofing

The earlier machine learning approaches to PAD were based on the extraction of handcrafted features such as Histograms of Oriented Gradient (HOG) [9], Differences of Gaussians (DoG) [10, 11], and Local Binary Patterns (LBP) [12, 13, 14]. Recently, deep learning has replaced traditional feature extraction, and the research focus has shifted towards the design of the most suitable neural network architectures. Yang *et al.* [15] were the first to use CNNs in face anti-spoofing, while, [16, 17], followed by [18, 19], proposed approaches competitive to the then state-of-the-art.

Better approaches were found to enhance results such as including Central Difference Convolutional Networks (CDCN), and transformers [1, 20, 21], and the use of a combination of more than one deep network type as in [22]. A newer approach is to rely on the use of independently trained neural networks to infer depth information [23, 24, 25], or Near Infrared (NIR) information [26]. Most recently, in this direction of work, [27] proposed the use of a dual-stream CNN framework. One stream uses learnable frequency filters to extract features in the frequency domain that are less influenced by variations in sensors and lighting, while the other stream uses standard RGB images to supplement these features. A hierarchical attention module is used to combine the information from these two streams at different stages of the CNN.

2.2 Anomaly detection in face anti-spoofing

In principle, applying anomaly detection to face anti-spoofing problems should lead to improved generalization capabilities, since it makes no assumptions about the type of attack or the environment in which it took place. Arashloo *et al.* [2] was the first to use anomaly detection for face anti-spoofing, using One-Class SRCs and One-Class SVMs as generative and non-generative classifiers respectively. One class GMMs were used in [28] and [29], while combinations of CNNs with one class classifiers were proposed in [30, 31, 32, 33]. Baweja *et al.* [34] introduced in the training a normally distributed pseudo-negative class and a pairwise confusion loss. Feng *et al.* [3] proposed a residual learning framework with a spoof cue generator and an auxiliary classifier. Abduh and Ivrissimtzis [35] used a convolutional autoencoder and augmented the training set with images from the in-the-wild.

2.3 Transformers

Transformers are for some years now the de facto standard in natural language processing (NLP) applications and recently have been established as a state-of-the-art computational technique in many computer vision problems too. The potential of the transformers in computer vision tasks was demonstrated by the groundbreaking Vision Transformer (ViT) [4], which is still in wide use, and it is the network that we use here. Liu *et al.* [36], introduced the Hierarchical Vision Transformer Using Shifted Windows, showing that it works very well as a general-purpose backbone for computer vision problems.

Regarding the use of transformers for anomaly detection in computer vision tasks, Mishra *et al.* [37] proposed a transformer based network for detecting and locating anomalous regions in images. By incorporating transformers, their method is sensitive to the spatial details of the patches, which are analyzed by a Gaussian mixture density network to identify anomalous regions. Lee *et al.* [38] proposed AnoVit, a ViT-based encoder-decoder for anomaly detection and localization, while Mukherjee *et al.* [39] proposed OCFormer, a one-class transformer for image classification.

Regarding the use of transformers in face anti-spoofing, George and Marcel [6] proposed a ViT-based model for the zero-shot PAD. Wang *et al.* [20] cross-layer

relation-aware attentions (CRA) and hierarchical feature fusion (HFF). Liu and Liang [40] proposed the Modality-Agnostic ViT (MA-ViT), using early fusion to aggregate data from all training modalities, improving the model's performance on arbitrary modal attacks. Finally, Huang *et al.* [41] use an adaptive transformer model for few-shot PAD across various databases. To the best of our knowledge, there is no study of transformer-based anomaly detection techniques for PAD.

3 THE ANOFORMER

The proposed method uses feature vectors provided by the pre-trained ViT and ResNet[5], which are then processed by a one-class classification technique. In our experimental study in Section 4 we show results obtained by the use of isolation forests and one-class SVMs. However, our focus is on training a decoder of the feature vectors and then comparing the reconstruction error against a threshold to take the classification decision.

3.1 Architecture

The architecture of the Anoformer is illustrated in Fig. 1. The backbone networks are ViT, which, has already demonstrated its potential as an embedding extractor for the face PAD problem in [6] where a two-class training of the ViT feature vectors gave results competitive to the state-of-the-art, and ResNet-18, both pre-trained on ImageNet [42].

Regarding the choice of specific ViT architecture, we first note that our proposed model can work with any version of ViT, providing compatibility with future improvements to ViT. Here, we employed the Data-Efficient Image Transformer [43] (DeiT-Base), which is an improved version of ViT of lightweight design. To learn diverse features, the training dataset's bona fide images are fed into the ViT and ResNet networks. The ViT divides the input image into patches and uses the extracted features as the sequence input for the transformer, followed by transformer layers. The transformer encoder layer is composed of multiple encoder blocks, each with multi-head self-attention (MSA) and multi-layer perceptron (MLP), as in [4].

The image patches undergo linear transformation to produce the queries (Q), keys (K), and values (V) of the self-attention mechanism, with position encoding (PE) added to keep track of each input token's position. The MLP contains two linear layers with a GELU activation function. Finally, the encoded patches are reshaped and projected into a reconstruction vector via a learned projection matrix.

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V \quad (1)$$

The ViT leverages the attention mechanism which gathers information from the entire input sequence. Self-attention layers scan through a sequence of elements and update them based on the information obtained from the entire sequence. In essence, they simulate explicitly every pair-wise interaction that occurs between the components of the input sequence. Thus, the self-attention maps, which are learned separately for each layer, are necessary for a transformer model to encode the dependencies between input tokens.

Fig. 2 shows attention maps from different ViT layers for a bona fide and an imposter input. We note the difference in ViT's behaviour between the two classes, in that certain prominent facial features such as eyes, nose, and mouth are more prominent in the imposter attention maps.

The decoder decodes the 768×1 reconstruction vector back to the original image shape. We used 4 transposed convolutional layers, with ReLU in between, except for the last layer, which is followed by a sigmoid as the final activation function. The decoder part of the Anoformer was trained with the features of the bona fide images extracted from ViT and ResNet. The decoder is trained with the objective of minimizing the error between the input and the output of the network, aiming at reconstructing bona fide images with high fidelity.

3.2 Implementation

To avoid contributions from the input image's background, we use the MTCNN algorithm for face detection, and cropped the images, retaining the face regions only. Then the images are rotated to have the eye centres horizontally aligned and finally resized to 224×224, which is the native resolution of the ViT transformer pre-trained on ImageNet.

In the final binary classification section, we used an Adam optimizer with initial learning rate of 1e-4 and batch size 16. A label smoothing cross entropy loss function was used to train the classifier. The MLP head is the binary classifier which contains two fully-connected layers of dimensions 512 and 2. The development environment was the PyTorch running on a PC with an Intel CPU, 64 GB of RAM, and Google Colab GPU.

Our backbone consists of two parts, the first part is the Deit-Base [43] which spatial position embeddings, to improve image processing capabilities which has an output embedding dimension of 768. The second component is a ResNet-18 network with 18 layers that have once more been trained on ImageNet. The size of the output from the ResNet is 2048. Therefore, in order to bring the size of the outputs (features) produced by ResNet down to the same level as those produced by transformer 768, we add one dense layer at the very end

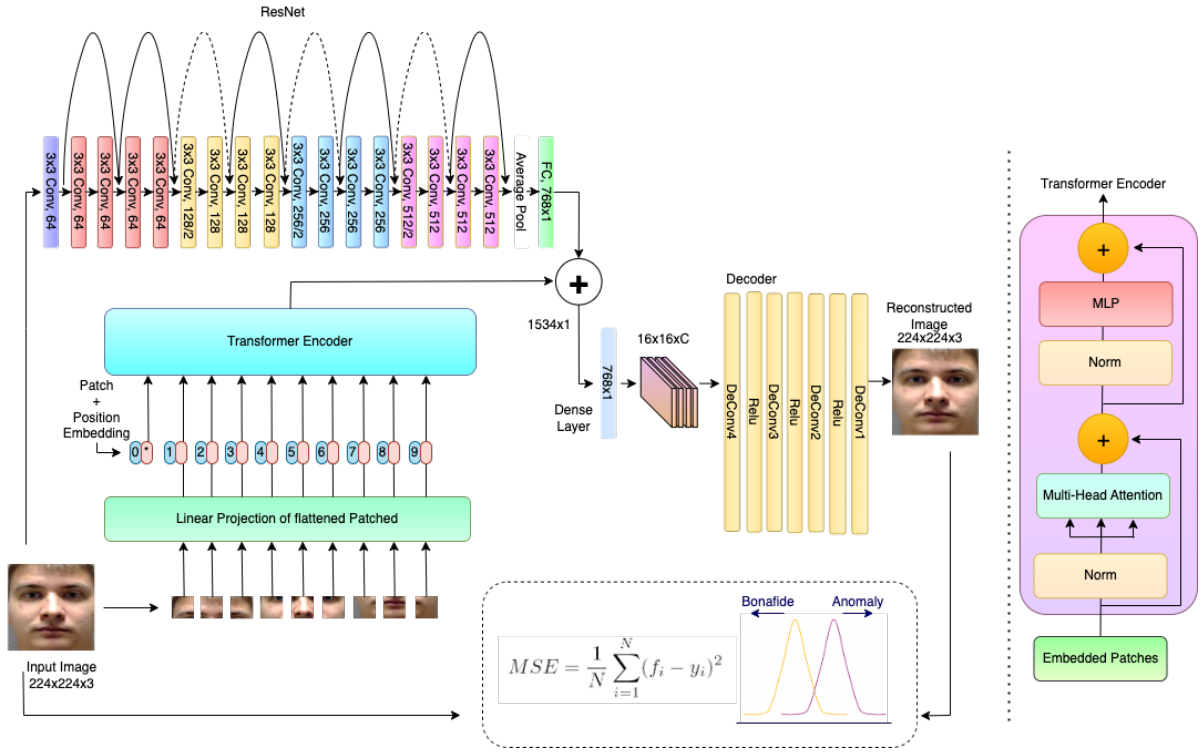

Figure 1: Architecture of the Anoformer.

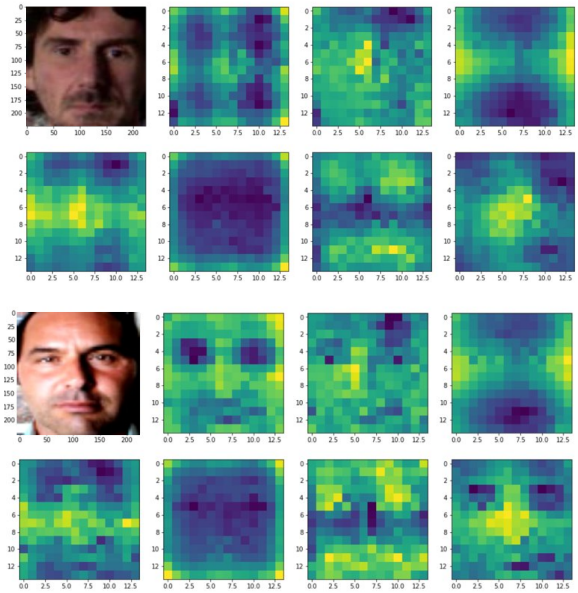

Figure 2: Visualisation of input faces from the Replay Attack database, and the attention maps of several ViT layers. **Top two rows:** a bona fide face. **Bottom two rows:** an imposter face.

of the network. Finally, then concatenate the two feature sets, the one from Deit-Base and that from Resnet-18, to create a vector of 1536 features. This data is then compressed to a size of 768 by adding one dense layer before being sent to the decoder.

The decoder was trained under the Mean Squared Error (MSE) loss function. While it is a pixel-level loss, assuming independence between pixels, it has been repeatedly shown that it works very well in practice, and its simplicity and the fact that it was supported by our development environment led to fast training times.

Regarding the computation of the anomaly score, that is, the error between the input and the reconstructed image, the natural choice is to use the loss function itself, and thus, we use MSE as our default. We experimented with other error metrics, such as Cosine Similarity, the Structural Similarity Index (SSIM), and the Frechet Inception Distance (FID) score [44]. We found that FID performed comparably to MSE, even though the decoder was trained with MSE, and thus, we include some relevant results in Section 4.

4 RESULTS

4.1 Databases

Our experiments were performed on two commonly used face anti-spoofing databases, the *Replay-Attack* (RA) [13], and the *Spoof in the Wild* (SiW) [17]. RA is a low-resolution dataset, containing live and spoof videos from 50 subjects, comprising three different presentation attack species, while SiW is a high-resolution dataset with live and spoof videos from 165 subjects, comprising eight different presentation attack species. The larger number of subjects, the larger number of

presentation attack species, and the higher variability in subject poses and lighting conditions, mean that SiW poses a more challenging classification problem to tackle. We also note that another advantage of SiW over other publicly available anti-spoofing datasets is its racial variety, including a sufficient number of African, Asian, Caucasian, and Indian subjects. It also has a good split between male and female subjects.

We divided the RA and SiW databases into training, validation, and testing sets with non-overlapping subjects. In the validation and testing datasets of the RA database, we included all three presentation attack species; printed photo, video, and digital photo. In the training and validation datasets of the SiW database, we included three representative attack species; printed photo, replay attack using iPhone, and replay attack using a tablet.

4.2 Evaluation Metrics

We report our results using the APCER, BPCER and ACER metrics, which are the most commonly used error metrics in face anti-spoofing, recommended by the ISO/IEC 30107-3:2023 [45] protocol for testing and reporting on biometric PAD.

The Attack Presentation Classification Error Rate (APCER) measures the performance of the system on attack images, that is, its ability to identify correctly spoof images. Unlike the most commonly used in binary classification problems False Positive Rate (FPR), to compute the (APCER), we compute misclassification rates separately over each attack species, and take the maximum. That is, APCER measures the system's performance under the most challenging type of attack, rather than under the average attack. The bona fide classification error rate (BPCER) is the misclassification rate over the bona fide samples. Finally, the ACER, which is considered a good measure of the overall performance of an algorithm, is just their average ACER=(APCER+BPCER)/2.

The definition of APCER as the maximum of the misclassification rates that are computed separately over each attack species brings to the fore an important methodological problem. When we measure the misclassification rates, do we use a single threshold for all attack species, or do we choose a different threshold for each one of them?

For example, in [6] different thresholds were used, and thus different BPCERs are reported as corresponding to each attack species, even though the dataset of bona fide presentations is one. Here, we use a single threshold for all attacks, firstly because in practice it is unrealistic to expect prior knowledge of the attack species that will inform the choice of threshold, and secondly, because we think it is closer to the spirit of the ISO definition of APCER, that is, to consider the worst case outcome over all attack species, rather than splitting the problem into smaller, easier to tackle sub-problems. In particular, we used the threshold corresponding to the Equal Error Rate (EER) on an independent validation set from the same database as the testing set.

4.3 Anoformer validation

In Tables 1 and 2 we report the results for four different classifiers over the ViT+ResNet backbone, tested on the RA and SiW databases, respectively. The one-class SVM (OC-SVM) is a widely used one-class classification method, being essentially an SVM trained with positively labelled data only, and aiming at maximising the separation of their class from the origin of the coordinate system. The second classifier we used is the Isolation Forest, which is based on decision trees and it is theoretically justified under the assumption that anomalies are "few and different". We note that while this is a very realistic assumption for the face anti-spoofing problem, it is not reflected in the usual PAD evaluation protocols that we also use here. Finally, in the last two rows of the tables, we report error rates for the Anoformer and the Anoformer with the FID metric for the computation of the anomaly score as discussed in Section 3. We notice that the combination of Anoformer with MSE in the reconstruction gives lower ACERs on both databases, and it is the configuration that we will evaluate.

Table 1: ViT + ResNet backbone with various one-class classifiers tested on RA

	ACER	APCER	BPCER
OC-SVM	.31	.26	.36
Isolation Forest	.33	.27	.40
Anoformer MSE	**.13**	**.23**	**.03**
Anoformer FID	.19	**.23**	.16

Table 2: ViT + ResNet backbone with various one-class classifiers tested on SiW.

	ACER	APCER	BPCER
OC-SVM	.31	.16	.46
Isolation Forest	.33	**.11**	.55
Anoformer MSE	**.21**	.33	**.10**
Anoformer FID	.22	.35	**.10**

Table 3 shows the results of the ablation study on the backbone of the Anoformer. The ViT + ResNet combination gives on both databases lower ACERs than ViT or ResNet alone, and notably the APCERs and BPCERs are both lower in both cases.

4.4 Performance evaluation

In Table 4, we compare the error rates of the proposed Anoformer against [34], which is a recently published

Table 3: Ablation study for the Anoformer backbone.

	RA			SiW		
	ACER	BPCER	APCER	ACER	BPCER	APCER
ViT	.16	.07	.25	.38	.42	.34
Res	.19	.07	.31	.44	.36	.52
V+R	**.13**	**.03**	**.23**	**.21**	**.10**	**.33**

anomaly detection method that reported results on the same databases and with the same error metrics as us. The results show that the Anoformer gives a lower ACER on both RA and SiW.

Table 4: Performance comparison against [34]

		ACER	APCER	BPCER
RA	[34]	.21	.25	.17
RA	ours	**.13**	**.23**	**.03**
SiW	[34]	.23	**.23**	.23
SiW	ours	**.21**	.33	**.10**

Finally, in Table 5, we report cross-database testing results for the Anoformer with threshold-specific metrics. As expected cross-database testing gives significantly higher ACERs. However, we note that this is mostly due to the higher APCERs.

Table 5: Intra- and cross-database testing of the Anoformer with threshold-specific metrics.

	ACER	BPCER	APCER
RA/RA	.13	.03	.23
SiW/RA	.26	.03	.50
RA/SiW	.27	.13	.42
SiW/SiW	.21	.10	.33

5 CONCLUSION

We proposed Anoformer, an anomaly detection model for PAD, with the pre-trained transformer ViT and the deep CNN Resnet in the backbone, and a one-class trained convolutional decoder for reconstruction. Our experimental results show that the performance of the model is competitive with the current state of the art in anomaly detection for generalised face anti-spoofing.

In the future we would like to test the Anoformer on more databases, and even use test bona fide image from outside the specialised PAD databases. Our aim would be to further demonstrate the generalisation power of anomaly detection.

6 REFERENCES

[1] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. CVPR*, 2020. [Online]. Available: 10.1109/CVPR42600.2020.00534

[2] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13 868–13 882, 2017.

[3] H. Feng, Z. Hong, H. Yue, Y. Chen, K. Wang, J. Han, J. Liu, and E. Ding, "Learning generalized spoof cues for face anti-spoofing," *arXiv:2005.03922*, 2020.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778. [Online]. Available: 10.1109/CVPR.2016.90

[6] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *Proc. IJCB*, 2021, pp. 1–8. [Online]. Available: 10.1109/IJCB52358.2021.9484333

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[8] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv:1811.12231*, 2018.

[9] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using hog–ebgm," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1537–1543, 2008.

[10] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bi-linear discriminative model," in *Proc. ECCV*, 2010, pp. 504–517.

[11] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *Proc. ICIP*, 2011, pp. 3557–3560.

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

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

[14] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analy-

sis," in *Proc. ICIP*, 2015, pp. 2636–2640.

[15] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv:1408.5601*, 2014.

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

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

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

[19] C. Nagpal and S. R. Dubey, "A performance evaluation of convolutional neural networks for face anti spoofing," in *Proc. IJCNN*, 2019, pp. 1–8.

[20] Z. Wang, Q. Wang, W. Deng, and G. Guo, "Face anti-spoofing using transformers with relation-aware mechanism," *IEEE Trans. BBIS*, vol. 4, no. 3, pp. 439–450, 2022. [Online]. Available: 10.1109/TBIOM.2022.3184500

[21] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "Nas-fas: Static-dynamic central difference network search for face anti-spoofing," *IEEE Trans. PAMI*, vol. 43, no. 9, pp. 3005–3023, 2020.

[22] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, "Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing," *IEEE Trans. IFS*, vol. 16, pp. 937–951, 2020.

[23] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *Proc. ECCV*, 2020, pp. 641–657.

[24] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," *IEEE Trans. BBIS*, vol. 3, no. 3, pp. 285–295, 2021.

[25] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proc. CVPR*, 2020, pp. 5042–5051. [Online]. Available: 10.1109/CVPR42600.2020.00509

[26] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Trans. IFS*, vol. 16, pp. 2759–2772, 2021.

[27] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper, "Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection," in *Proc. WCACV*, 2022, pp. 3722–3731.

[28] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing," in *Proc. ICB*, 2018, pp. 75–81. [Online]. Available: 10.1109/ICB2018.2018.00022

[29] F. Xiong and W. AbdAlmageed, "Unknown presentation attack detection with face rgb images," in *Proc. BTAS*, 2018, pp. 1–9.

[30] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, "Deep anomaly detection for generalized face anti-spoofing," in *Proc. CVPR*, 2019, pp. 0–0.

[31] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, "Spoofing attack detection by anomaly detection," in *Proc. ICASSP*, 2019, pp. 8464–8468.

[32] S. R. Arashloo, "Unseen face presentation attack detection using sparse multiple kernel fisher null-space," *IEEE Trans. CSVT*, vol. 31, no. 10, pp. 4084–4095, 2020.

[33] S. R. Arshloo, "Matrix-regularized one-class multiple kernel learning for unseen face presentation attack detection," *IEEE Trans. IFS*, vol. 16, pp. 4635–4647, 2021.

[34] Y. Baweja, P. Oza, P. Perera, and V. M. Patel, "Anomaly detection-based unknown face presentation attack detection," in *Proc. IJCB*, 2020, pp. 1–9.

[35] L. Abduh and I. Ivrissimtzis, "Training dataset construction for anomaly detection in face anti-spoofing," in *Proc. CGVC*, 2021.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV*, 2021, pp. 10 012–10 022.

[37] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *Proc. ISIE*, 2021, pp. 01–06.

[38] Y. Lee and P. Kang, "Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder," *IEEE Access*, vol. 10, pp. 46 717–46 724, 2022.

[39] P. Mukherjee, C. K. Roy, and S. K. Roy, "Ocformer: One-class transformer network for image classification," *arXiv:2204.11449*, 2022.

[40] A. Liu and Y. Liang, "Ma-vit: Modality-agnostic vision transformers for face anti-spoofing," in *Proc. IJCAI*, 2022, pp. 1180–1186.

[41] H.-P. Huang, D. Sun, Y. Liu, W.-S. Chu, T. Xiao, J. Yuan, H. Adam, and M.-H. Yang, "Adaptive transformers for robust few-shot cross-domain face anti-spoofing," *arXiv:2203.12175*, 2022.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical

image database," in *Proc. CVPR*, 2009, pp. 248–255.

[43] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*. PMLR, 2021, pp. 10 347–10 357.

[44] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[45] Biometrics IJS, "Iso/iec 30107-3: 2023. information technology - biometric presentation attack detection - part 3: Testing and reporting," *International Organization for Standardization*, 2023.