# Two-Stage Face Detection and Anti-spoofing

M. Faisal Nurnoby[1] and El-Sayed M. El-Alfy[1,2(✉)]

[1] Information and Computer Science Department, College of Computing and Mathematics, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
alfy@kfupm.edu.sa

[2] Fellow SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Interdisciplinary Research Center of Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

**Abstract.** Face recognition is a widely used biometric technique that has received a lot of attention. It is used to establish and verify the user's identity, and subsequently grant access for authorized users to restricted places and electronic devices. However, one of the challenges is face spoofing or presentation attack allowing fraudsters who attempt to impersonate a targeted victim by fabricating his/her facial biometric data, e.g., by presenting a photograph, a video, or a mask of the targeted person. Several approaches have been proposed to counteract face spoofing known as face anti-spoofing techniques. This paper's major goals are to examine pertinent literature, and develop and evaluate a two-stage approach for face detection and anti-spoofing. In the first stage, a multi-task cascaded convolutional neural network is used to detect the face region, and in the second stage, a multi-head attention-based transformer is used to detect spoofed faces. On two benchmarking datasets, a number of experiments are carried out and examined to assess the proposed solution. The results are encouraging, with a very high accuracy, which encourages further research in this direction to build more robust face authentication systems.

**Keywords:** Presentation attack · Biometric authentication · Face recognition · Face anti-spoofing · Deep learning · Vision Transformer

## 1 Introduction

Biometric security systems offer a plethora of convenient options for access control and surveillance. One of the most common and frequently utilized biometric technologies today is face recognition, which is increasingly built into a variety of devices for various security needs. Additionally, it is more convenient, non-intrusive, and efficient relative to other biometric traits [12]. It has attracted growing interest and become among the most active research areas of computer


G. Bebis et al. (Eds.): ISVC 2023, LNCS 14361, pp. 445–455, 2023.
https://doi.org/10.1007/978-3-031-47969-4_35

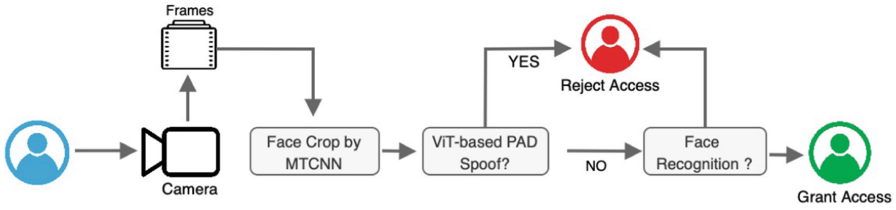



**Fig. 1.** Presentation Attack Detection System

vision, with numerous applications and commercial systems for access control, authentication of mobile payments, surveillance, human-computer interactions, and forensics. However, most of the face recognition systems are susceptible to assaults like spoofing, also called a face presentation attack [6]. A spoofing attack is a security threat that occurs when someone attempts to bypass a face recognition biometric system by presenting a fake face in front of the camera. A facial image can be simply put on photo paper, or printed on various masks, such as resin and silicon masks, as part of the presentation attack. Sometimes it can even be displayed on very high-resolution hand-held devices, including tablets, mobile phones, and laptops. The faces that are shown can be presented in front of any access control system [14]. Due to the widespread usage of devices that require this type of access, presentation or spoofing attacks have become increasingly sophisticated, posing considerable difficulties for the security of authentication systems.

An insecure face recognition system is always vulnerable to presentation attacks. A presentation attack detection (PAD) system aims to detect and prevent these kinds of attacks. Figure 1 shows an authentication system with PAD. Both impersonation and identity obfuscation are used in presentation spoofing. In an impersonation spoof attack, the attacker attempts to bypass authentication as someone else; in contrast, obfuscation attacks aim to conceal the identity of the attacker [10]. Any object used as a tool for presentation attacks is called spoof presentation assault instrument (PAI) [12]. PAIs include photo printouts, video replays, or 3D masks. In order to detect attacks such as replay and 2D printouts, features including color, texture, motion, and physiological indications, as well as CNN-based techniques, are frequently applied in the literature. Most presentation attack detection techniques found in current research focus on a small range of PAIs, and they are mostly images in the visible spectrum. Any PAD module witnesses a remarkable gain in performance after applying recent deep network models. Yet those methods do not generalize well to more realistic contexts [14]. Several cues can help detect the liveness of a face image and contribute to performance gains. Usually, those cues are able to prevent a particular type of attack, if not all [3]. Convolutional neural network (CNN)-based models are showing comparatively better performances. Nevertheless, those models are mostly 2D and are possibly vulnerable to attacks such as partial and 3D. As a remedy, many researchers moved their focus to the multi-channel approach (such as [17]) to trade accuracy with costs in terms of time and hardware.

In the area of computer vision, the application of self-attention based transformer architecture is still limited, even though it is a state-of-the-art tool for natural language processing tasks. The attention layer is either used in combination with convolutional layers in vision or is used to replace some layers of the convolutional network's constituent parts while maintaining the overall structure of the network. A pure transformer applied directly to sequences of picture patches may successfully perform image classification tasks without the need for CNNs. Besides, transformers are able to perform tasks on test data at a much faster speed compared to vanilla CNN models due to matrix operations and the efficient use of GPUs. Moreover, we aspire to investigate the effectiveness of a PAD when random-patched face images are fed to a ViT-based model. Our hypothesis is that detecting complex presentation attacks with a ViT [8] might bring about a more efficient solution.

Our work employed a fine-tuned Vision Transformer based model for face presentation attack detection. We argue that most of the existing solutions in the area of PAD show overfitting on the training dataset, and do not generalize well towards new and unseen spoofing attack samples. However, due to the patch-based nature of the ViT architecture, it showed more robust behavior when presented with new attacks or benign samples. We performed an ablation study through a series of experiments in order to measure the impacts of different environmental components in the experiments.

## 2 Related Works

Several approaches have been proposed in the literature for presentation attack detection (PAD). Moreover, a number of datasets have been introduced in the literature for benchmarking such as Replay-Attack, Replay-Mobile, SWAX, NUAA, SiW-M, CASIA-FASD, MSU-MFSD, CASIA-SURF, HiFiMask, WMCA, CelebA-Spoof and OULU-NPU. The majority of the research focus on face PAD, particularly 2D printouts and replay attacks. In this section, we provide a quick overview of recent PAD approaches.

### 2.1 Feature-Based PAD Systems

One of key findings is that the majority of prior face anti-spoofing research relied on a classifier based on hand-crafted features like motion patterns [2], image quality [9], Local Binary Patterns (LBP) [5], and image distortion measure [21]. It was recommended in [21] to use an image distortion analysis (IDA) face spoof detection method. Four different features, including specular reflection, blurriness. chromatic moment, and color diversity, are extracted to form the IDA feature vector. To differentiate between real (live) and spoof faces, an ensemble classifier made up of various SVM classifiers trained for various face spoof attempts (such as printed images and replayed videos) is employed. The suggested method is expanded to include voting-based multiframe face spoof detection in videos. According to the authors of [22], texture characteristics like

LBP, DoG, or HOG may be employed to differentiate between fake faces and actual faces. Texture-based techniques have been quite successful in the Idiap and CASIA databases.

### 2.2 CNN-Based PAD Systems

Despite the possibility of a wide range of PA, most of the works in the literature have detected 2D-type attacks such as replays and prints, primarily because such PAIs are simple to produce. However, the majority of the most recent cutting-edge findings come from techniques based on CNN.

The authors in [14] extended their previous work [13] on a deep patch-based CNN model for face PAD. The article argued that a model trained with face image patches can detect spoofs better. One important takeaway from the experiment is that the patch-based approach helps the model avoid memorizing the background of a face image. As the proposed approach leverages image patches, it requires less image data.

In [18], the authors applied a meta-teacher architecture with pixel-wise supervision in order to oversee presentation attack detectors. They demonstrated that, in comparison to teacher-student models and hand-crafted labels, the metateacher model offers better-suited monitoring. They found that the meta-teacher approach offered adequate supervision without the need for pixel-wise class labels when training networks. The article [19], which formulates the PAD issue as problems of both zero-shot learning and few-shot learning. They used live and spoof films, together with a few examples of novel assaults, for training a fine-tuned meta-learner that focuses on identifying hidden face spoof attack types. They demonstrated how their approach can outperform the competition on PAD benchmarks that are already in use.

For recognizing presentation attacks, the authors of [7] suggested a method, known as FaceSpoof Bluster, that combines deep neural networks and inherent visual features. The technique uses a pre-trained CNN model and an SVM classifier to analyze illumination, salience, and depth maps of face images in conjunction with to provide robust and discriminating features. Each of these attributes is identified separately, and then a meta-learning classifier combines them to produce better results.

### 2.3 Multichannel PAD Systems

The fundamental principle of multichannel approaches is that they use complimentary data acquired from many channels, which makes it very difficult for the adversaries to trick the spoof detection systems. According to the channels employed in the PAD system, an attacker must duplicate the characteristics of a genuine sample to a number of sensing domains; this makes the system harder to compromise [11].

George et al. in [12] argued systems that uses multiple channels, such as color, depth, infra-red, and thermal, might be more useful in addressing this presentation attack problem. They purposed a multi-channel Convolutional Neural

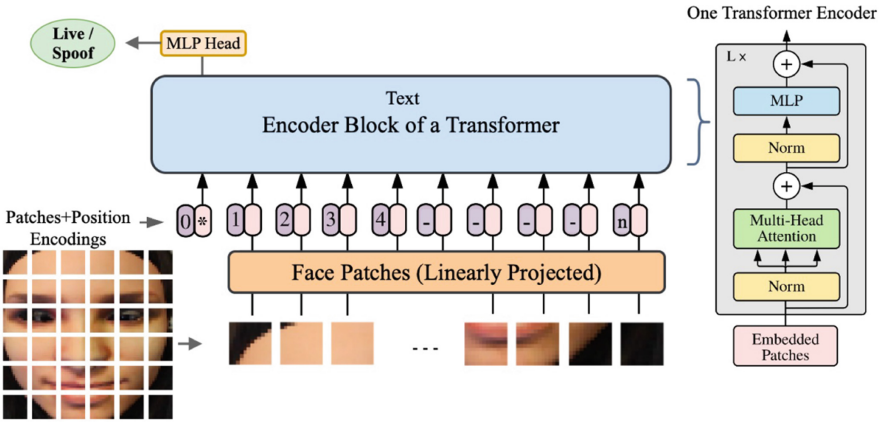



**Fig. 2.** Proposed face transformer based model

Network (MCCNN)-based PAD method. They also developed a new database for face PAD called Wide Multi-Channel Presentation Attack (WMCA), which comprises a wide range of 2D and 3D presentation assaults for both impersonation and obfuscation attacks. This database is publicly accessible through Idia Research Institute in Switzerland and can be found at[1] and contains 1941 short video recordings of 72 different identities. To enhance the study of facial PAD, data from many channels, including color, depth, near-infrared, and thermal, is present. The suggested methodology outperformed the baselines when compared to feature-based methods, attaining ACER score of 0.3% on the introduced dataset. In [10], the authors readdressed the same problem and presented a new framework to learn a robust PAD system. By adding a unique loss function that requires the network to learn a compact embedding for the legitimate class while being removed from the representation of attacks, the paper further expands the MCCNN technique to a one-class classifier framework. A unique method of learning a strong spoof attack detection system from real and known attack categories is introduced in the presented framework. The usefulness of the suggested method is demonstrated by the higher performance in unseen assault samples in the WMCA database.

Some recent papers on PAD have applied diverse approaches. The work in [16] outlined a video processing method named Temporal Sequence Sampling (TSS) by discarding affine movement and finally encoding the video into a single RGB image. The work applied a CNN with self-supervised learning for automatic labeling. Another approach for PAD based on CNN and background subtraction with ensemble classifiers is presented in [4]. Experiments were conducted for merely three attack types from four datasets in a selective manner. More recently, the authors in [1] used a unified DNN approach for PAD by integrating all the standalone tasks. They applied an ensemble approach with ViT-based features.

[1] https://paperswithcode.com/dataset/wmca.

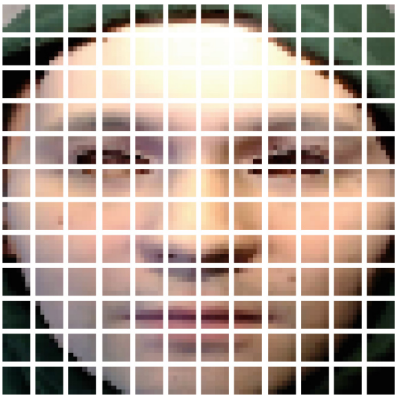


**Fig. 3.** Sample Face Patches from the SiW dataset

**Fig. 4.** Samples from the NUAA dataset

## 3 Methodology

In this study, we developed a transfer-learning model based on a vision transformer for detecting the face presentation attack. The workflow is explained in the following subsections.

### 3.1 Face Detection

The MTCNN (Multi-Task Cascaded Convolutional Neural Network) is used to recognize faces in the color channel. First, we extract image frames from the video files and take only the 20th frame of the sequence as the videos have very minor facial movements. We crop the faces during the preprocessing stage to remove the influence of the background. The MTCNN algorithm is then used for face detection and landmark localization. The eyes' centers are horizontally aligned on the chosen faces. The photos are then reduced to a resolution of 128 $\times$ 128 after this alignment to feed the transformer.

### 3.2 Anti-spoofing Model Architecture

In our work, we evaluated a vision transformer (ViT) as shown in Fig. 2 to detect presentation attacks on facial images, drawing inspiration from the success of the transformer application in NLP. We applied a pretrained vision transformer as proposed by Alexey Dosovitskiy et al. [8] as the backbone of our network. Without the use of convolution layers, the ViT model applies the Transformer architecture to a sequence of picture patches with self-attention. To handle 2D images, an image is transformed into a series of flattened 2D patches, which also act as the transformer's useful input sequence. The patches are flattened and mapped using a trainable linear projection known as the patch embeddings because the transformer employs a constant latent vector across all of its layers.

### 3.3 Model Implementation Details and Training

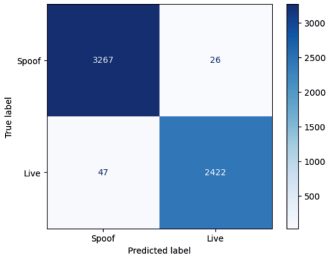
The vision transformer models surpass cutting-edge techniques in several vision benchmarks despite having been trained with a lot of data. A huge model must be completely retrained, which is computationally expensive. On the other hand, fine-tuning may make use of these potent models in settings with little data without necessitating a lot of computing capacity. We used a model trained on $6 \times 6$ patches; hence, the length of the input sequence is the same as the patches' number times THE height-width of the input image, which is 72 for both. Therefore, a face image would consist of 144 patches in total (Fig. 3). To keep positional information, a 1D positional embedding is additionally inserted in addition to patch embeddings. For the classification task, an MLP head with a fully connected layer was stacked on top of the transformer layers. To be more specific, we adjusted the model and applied binary cross-entropy loss. Then we replaced the last MLP layer and added a single output layer. We have examined the effect of fine-tuning.

During the training phase, we applied data augmentation using random horizontal flips. A fixed learning rate of 0.001 and a weight decay value of 0.0001 were used to supervise the network. During training, a batch size of 256 was employed. We trained the model on a GPU grid of a Macbook-M1-Pro/32GB-RAM computer for 25 epochs using the common Adam Optimizer. We chose our optimum model considering the lowest loss value in the validation dataset. Keras-Tensorflow was used to implement the architecture.

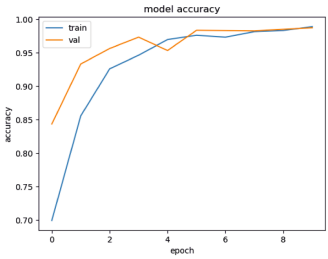
## 4 Experimental Work

### 4.1 Datasets and Performance Measures

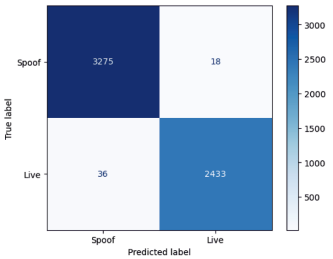
Two datasets have been used in our experiments: *Spoof in Wild with Multiple Attack (SiW-M) Database*: This dataset contains a large range of attacks recorded with an RGB camera [15]. It contains 493 subjects with 660 live samples and 968 attack samples. Altogether, they make 1,628 files with a total of 13



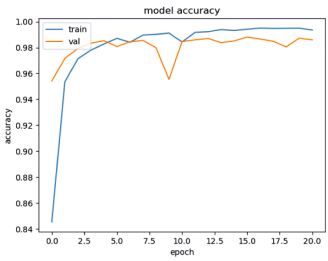

(a) CM: Vit-PadSiWAug



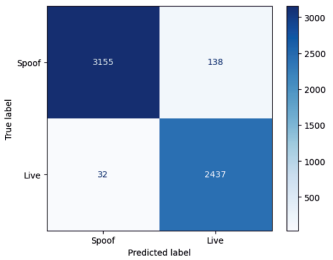

(b) Acc: Vit-PadSiWAug



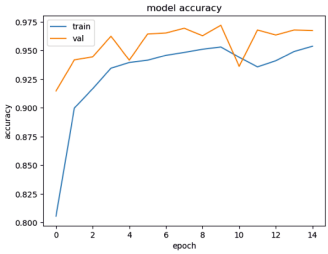

(c) CM: ResNet101PredPadSiW



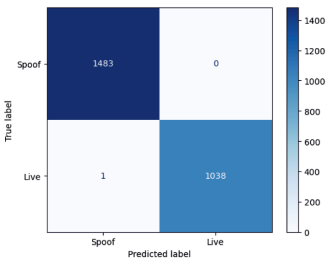

(d) Acc: ResNet101PredPadSiW



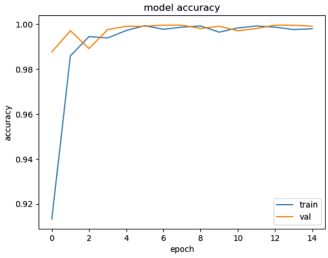

(e) CM: pretVgg19S




(f) Acc: pretVgg19S




(g) CM: VitPadNA




(h) Acc: VitPadNA

**Fig. 5.** The confusion matrix (CM) and accuracy (ACC) curve for each model

possible poses, expressions, and lighting variations. The attack samples were generated from multiple sessions. The attacks include a variety of makeups, masks, partials, and 2D presentation assaults.

*NUAA Photo Imposter Database*: This is a publicly available imposter database created by using a generic WebCam [20]. It has 15 subjects and images were captured in two sessions. Images of both the live subjects and their pictures were taken during each session. Each subject was instructed to look directly into the camera during the capture, maintaining a neutral expression and refraining from any outward motions, such as head movements or blinking eyes. Figure 4 shows some screenshots of the recorded photos.

As per ISO/IEC 30107-3 metrics, we used three performance measures in addition to the accuracy (ACC): Bonafide Presentation Classification Error Rate (BPCER), Attack Presentation Classification Error Rate (APCER), Average Classification Error Rate (ACER).

### 4.2 Experiments and Results

Three CNN-based baselines have been applied for comparison with the proposed ViT-based approach. We compared our model with other transfer learning-based models that are from two well-known architectures, such as VGG, and ResNet101, as the presented approach is also a transfer learning-based one.

We have conducted a set of six rounds of experiments with the SiW-M and NUAA datasets. In particular, we evaluated the proposed ViT-based PAD model on the two datasets and compared the model with two pretrained models dubbed as pretResNet101S and pretVGG19S. The outcomes of the six experiments are demonstrated in Table 1. The first experiment was on the SiW-M dataset with the proposed model ViTPadS without data augmentation during the training. The model shows a higher APCER score compared to BPCER. However, the same model trained with augmented data shows better scores. The pretrained baseline model, dubbed as pretVGG19PadS, scored lower accuracy, thus having higher scores in both APCER and BPCER. The best performance on the SiW-M dataset was shown by the pretrained ResNet101-based model with the lowest ACER score of 1.0. The experiment we carried out on the NUAA dataset using the proposed model, VitPadNA, scored an ACER score of 0.05. We also record the training and validation accuracy curves of our experiments. The confusion matrix and accuracy curves for each model are shown in Fig. 5. All curves show decent behaviors except for the experiment of the VitPadNA model which was trained and tested on the NUAA dataset. The model seems to converge quickly; it only required two epochs. This is due to the small data size. It is also too much biased for the training dataset. Also, the NUAA dataset has fewer variations in data.

**Table 1.** Comparison of various models using four performance measures

| Model | Dataset | APCER | BPCER | ACER | ACC |
|---|---|---|---|---|---|
| ViTPadS | SiW-M | 1.9 | 0.79 | 1.35 | 98.7 |
| ViTPadSA | SiW-M | 1.46 | 0.58 | 1.02 | 99.01 |
| pretResNet101S | SiW-M | 1.46 | 0.55 | 1.0 | 99.10 |
| pretVGG19PadS | SiW-M | 1.30 | 4.19 | 2.74 | 97.10 |
| ViTPadNA | NUAA | 0.01 | 0.0 | 0.05 | 99.60 |
| ViTPadMulti | SiW-M | – | – | – | 97.70 |

## 5 Conclusion and Future Work

Face recognition systems are one of the most prominent biometric tools for numerous access control solutions. This paper explores a two-stage approach for face detection and anti-spoofing. Faces are detected using a multi-task cascaded convolutional neural network, and a vision-transformer-based model, dubbed ViTPadS, is developed for face presentation attack detection. Two datasets were used to evaluate the performance of the model with or without data augmentation. The model's performance is also compared to two other pretrained models used as baseline models. The experimental work demonstrated promising results for the proposed ViT-based model. This work can be further improved by conducting an ablation study and fine-tuning. Also, to make the model more robust against any unknown attacks, cross-dataset testing with a zero or few-shot learning architecture can also be considered.

**Acknowledgment.** The authors would like to thank King Fahd University of Petroleum and Minerals for support under the Interdisciplinary Research Center for Intelligent Secure Systems Grant no. INSS2204.

## References

1. Al-Refai, R., Nandakumar, K.: A unified model for face matching and presentation attack detection using an ensemble of vision transformer features. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 662–671 (2023)
2. Anjos, A., Marcel, S.: Counter-measures to photo attacks in face recognition: a public database and a baseline. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–7 (2011)
3. Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face anti-spoofing using patch and depth-based CNNs. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 319–328 (2017)
4. Benlamoudi, A., et al.: Face presentation attack detection using deep background subtraction. Sensors **22**(10), 3760 (2022)
5. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: IEEE International Conference on Image Processing (ICIP), pp. 2636–2640 (2015)

6. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Trans. Inf. Forensics Secur. **11**(8), 1818–1830 (2016)
7. Bresan, R., Pinto, A., Rocha, A., Beluzo, C., Carvalho, T.: FaceSpoof buster: a presentation attack detector based on intrinsic image properties and deep learning. arXiv preprint arXiv:1902.02845 (2019)
8. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Galbally, J., Marcel, S., Fierrez, J.: Image quality assessment for fake biometric detection: application to iris, fingerprint, and face recognition. IEEE Trans. Image Process. **23**(2), 710–724 (2013)
10. George, A., Marcel, S.: Multi-channel face presentation attack detection using deep learning. In: Ratha, N.K., Patel, V.M., Chellappa, R. (eds.) Deep Learning-Based Face Analytics. ACVPR, pp. 269–304. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-74697-1_13
11. George, A., Marcel, S.: On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8 (2021)
12. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans. Inf. Forensics Secur. **15**, 42–55 (2019)
13. Kantarcı, A., Dertli, H., Ekenel, H.K.: Shuffled patch-wise supervision for presentation attack detection. In: IEEE International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5 (2021)
14. Kantarcı, A., Dertli, H., Ekenel, H.K.: Deep patch-wise supervision for presentation attack detection. IET Biomet. **11**(5), 396–406 (2022)
15. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4680–4689 (2019)
16. Muhammad, U., Yu, Z., Komulainen, J.: Self-supervised 2D face presentation attack detection via temporal sequence sampling. Pattern Recogn. Lett. **156**, 15–22 (2022)
17. Nikisins, O., George, A., Marcel, S.: Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In: IEEE International Conference on Biometrics (ICB), pp. 1–8 (2019)
18. Qin, Y., Yu, Z., Yan, L., Wang, Z., Zhao, C., Lei, Z.: Meta-teacher for face anti-spoofing. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 6311–6326 (2021)
19. Qin, Y., et al.: Learning meta model for zero-and few-shot face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11916–11923 (2020)
20. Tan, X., Li, Y., Liu, J., Jiang, L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 504–517. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15567-3_37
21. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Trans. Inf. Forensics Secur. **10**(4), 746–761 (2015)
22. Yang, J., Lei, Z., Liao, S., Li, S.Z.: Face liveness detection with component dependent descriptor. In: IEEE International Conference on Biometrics (ICB), pp. 1–6 (2013)