

On the Effectiveness of Vision Transformers for Zero-shot Face Anti-Spoofing

Anjith George and Sébastien Marcel
 Idiap Research Institute
 Rue Marconi 19, CH - 1920, Martigny, Switzerland
 {anjith.george, sebastien.marcel}@idiap.ch

Abstract

The vulnerability of face recognition systems to presentation attacks has limited their application in security-critical scenarios. Automatic methods of detecting such malicious attempts are essential for the safe use of facial recognition technology. Although various methods have been suggested for detecting such attacks, most of them over-fit the training set and fail in generalizing to unseen attacks and environments. In this work, we use transfer learning from the vision transformer model for the zero-shot anti-spoofing task. The effectiveness of the proposed approach is demonstrated through experiments in publicly available datasets. The proposed approach outperforms the state-of-the-art methods in the zero-shot protocols in the HQ-WMCA and SiW-M datasets by a large margin. Besides, the model achieves a significant boost in cross-database performance as well.

1. Introduction

Face recognition offers a simple yet convenient way for access control. Though face recognition systems have become ubiquitous [23], its vulnerability to presentation attacks (a.k.a spoofing attacks) [29], [22] limits the application of these systems in safety-critical applications. An unprotected face recognition (FR) system might be fooled by merely presenting artifacts like a photograph or video in front of the camera. The artifact used for such an attack is known as a presentation attack instrument (PAI).

As the name indicates, presentation attack detection (PAD) systems try to protect FR systems against such malicious attempts. Though a wide variety of presentation attacks are possible, the majority of the research efforts have focussed on the detection of 2D attacks such as prints and replays, mainly due to the easiness of producing such attack instruments. Most of the research in PAD focus on the detection of these attacks using the RGB spectrum alone, using either feature-based methods or Convolutional Neural Network (CNN) based approaches. Several feature-based methods using color, texture, motion, liveliness cues, his-

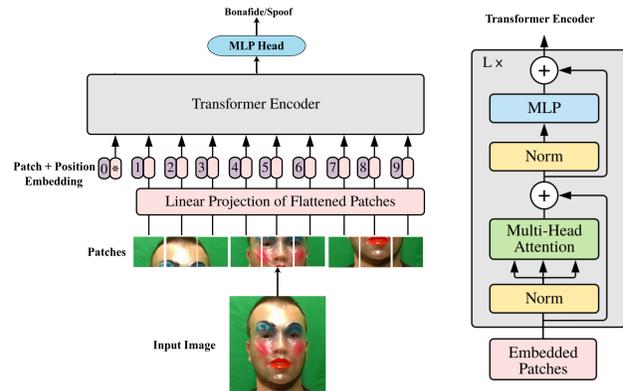


Figure 1. Vision Transformer model adapted for the presentation attack detection (PAD) task. The final layer is replaced and fine-tuned for the binary classification task (Image adapted from [11] and [36]).

togram features [6], local binary pattern [28], [10] and motion patterns [2] have seen proposed over the years for performing PAD. However, most of the recent state-of-the-art results are from CNN-based methods. Specifically, CNNs using auxiliary information in the form of binary or depth supervision have shown to improve performance greatly [3, 14]. Nevertheless, the majority of these methods perform well only in the case of 2D attacks and the performance of such methods degrade when evaluated against sophisticated 3D and partial attacks [27]. Even in the case of 2D attacks, these models often fail to generalize towards unseen attacks and environments.

Recently, several multi-channel approaches have been proposed to address the limitations of PAD systems [18, 31, 17]. Though such methods achieve superior performance as compared to RGB methods, the cost of additional hardware required limits the application of such methods to protect legacy RGB-based FR systems. Hence, it is desirable to have a robust RGB-based PAD method that is robust against a wide variety of 2D, 3D, and partial attacks. Ideally, a PAD system should be able to generalize well to unseen attacks and environments.

In most cases, the amount of data available to train a PAD model is very limited. This limits the possibility of training

deep architectures from scratch. And from the literature, transfer learning from a pre-trained network has proven to be an effective strategy to deal with the limited data problem. Moreover, a pre-trained network could also help with addressing the domain shift as the model has seen a wide variety of images in different environments.

In this work, we investigate the effectiveness of the vision transformer model [11] for the zero-shot presentation attack detection problem. We compare the proposed method with both state-of-the-art methods and fine-tuned CNN models. Specifically, we investigate the performance of this method in challenging unseen attack and cross-database scenarios.

To the best knowledge of the authors, this is the first work using vision transformers for the presentation attack detection task. The main contributions of this work are listed below.

- Introduces a simple yet effective Vision Transformer-based PAD framework.
- Shows the effectiveness of the vision transformer framework in an unrelated downstream task while adapting only a minimal number of parameters in the training stage.
- The proposed approach has been extensively evaluated in challenging unseen and cross-database conditions and it achieves the state of the art performance, outperforming other baselines by a large margin.

Additionally, the source code and protocols to reproduce the results are available publicly¹ to allow further extension of the work.

2. Related work

Presentation Attack Detection: The majority of face presentation attack detection methods deal with the detection of 2D attacks. And most of these methods capitalize on capturing the quality degradation of the samples during recapture. Feature-based methods such as motion patterns [2], Local Binary Patterns (LBP) [6], image quality features [13], and image distortion analysis [37] have been utilized over the years for PAD. Most of the recent PAD methods use CNN based approaches [26, 14, 3]. Since many of these methods depend on quality degradation during recapture to distinguish attacks from bonafide, they may not be suitable for the detection of sophisticated attacks like 3D masks and partial attacks. Since the quality of attack instruments improves over time, the PAD methods used should be robust against unseen attacks as well.

Multi-channel methods have been proposed in the literature as a solution to handle a wide variety of attacks

[33, 34, 33, 12, 5]. The core idea of multi-channel/spectral methods is the usage of complementary information from different channels making it extremely hard for attackers to fool the PAD systems. An attacker would have to replicate the properties of a bonafide sample across different sensing domains, making it difficult depending on the channels used in the PAD system. In [18], George *et al.* presented a multi-channel face presentation attack detection framework using color, depth, infrared and thermal channels. Several recent works have also achieved good PAD performance utilizing multi-channel information [20, 16, 15].

Though multi-channel information could alleviate the issues with the PAD systems, the cost of the hardware increases with additional channels used. Moreover, it limits the widespread adoption of the PAD system, leaving legacy face recognition systems unprotected. An RGB-only PAD framework, which performs robustly across unseen attacks and environments is essential for the secure usage of legacy face recognition systems.

Transformer Models: The Transformer models proposed in [36] introduced a novel approach towards sequence transduction tasks obviating the need for convolution or recurrence mechanisms. Transformers essentially capitalize on the attention mechanism to model dependencies between input and output. As opposed to other recurrent methods, Transformer allows significant parallelization in sequence tasks, while achieving the state of the art results in many tasks. There have been several attempts to use the Transformer framework for vision tasks, some of the methods used initial layers of pre-trained models [25] to get features to be used in the Transformer model. Various forms of position encodings, i.e., fixed or learned are also added to the features before passing to the Transformer layers. The work in [8] used the encoder-decoder structure of the Transformer for the object detection task.

3. Proposed method

In this work, we propose to use transfer learning from a pre-trained Vision Transformer model for zero-shot face anti-spoofing task. The different stages of the framework are detailed in the following sections.

3.1. Preprocessing

The CNN model accepts images of resolution 224×224 . To avoid the contribution from the background and other database biases, we crop the face regions in the preprocessing stage. First, face detection and landmark localization are performed using the MTCNN [39] algorithm. The detected faces are aligned so that the eye centers are horizontally aligned. After this alignment, the images are cropped to a resolution of 224×224 .

¹Source code: https://gitlab.idiap.ch/bob/bob.paper.ijcb2021_vision_transformer_pad

3.2. Network Architecture

The proposed framework uses the recently proposed vision transformer [11] architecture as its back-bone. The details are given in the following sections.

The Vision Transformer model: Transformers were initially proposed by Vaswani *et al.* [36] for machine translation applications. These models leverage the attention mechanism and this framework has found applications in many natural language, audio, and vision tasks. The attention layers [4] aggregate information from the entire length of the input sequence. Transformers introduced self-attention layers that scan through and update each element in a sequence using the information from the whole sequence. Essentially, they explicitly model all the pairwise interactions between the components in the input sequence. Recently authors in [11] applied the standard transformer with minimal modifications for the image classification task. An image is divided into patches, and embeddings obtained from the patches are used as the sequence input for the transformer. The vision transformers introduce a new way for image classification instead of using convolutional layers. A sequence of image patches is used as the input followed by transformer layers.

While trained with large amounts of data, the vision transformer models outperform the state-of-the-art methods in many vision benchmarks. However, retraining such a large model from scratch is very computationally expensive. However, fine-tuning offers a way to utilize these powerful models in limited data scenarios without requiring much computational power. We used the model trained with 16×16 patches, meaning the input sequence length will be the number of patches $\frac{HW}{16^2}$, where H and W are the height and width of the input image in pixels. In addition to patch embeddings, a 1D positional embedding is also added to retain positional information. After the transformer layers, an MLP head consisting of a fully connected layer was added for the classification task.

In this work, we investigate the transferability of the pre-trained Vision Transformer model for the PAD task. Specifically, we replace the last layer with a fully connected layer with one output node and we fine-tune the model using binary cross-entropy loss (BCE). We have conducted experiments with adapting a different set of layers to find the effect of fine-tuning when trained with a small dataset. The framework used for PAD is depicted in Fig. 1.

Implementation details: We adapted the Vision Transformer base network described in [11], from the open-source implementation provided in [38]. Pretrained weights provided was used to initialize the network. Specifically, we used the “base” variant of the pre-trained model, made available for an image size of 224×224 , with a patch size of 16×16 . We used the model with the same resolution

as other baselines (224×224), to enable ready comparison between other ImageNet pre-trained models. Data augmentation was performed during the training phase with random horizontal flips with a probability of 0.5. The network was supervised with binary cross-entropy loss (BCE), with a fixed learning rate of 1×10^{-4} and a weight decay parameter of 1×10^{-5} . A batch size of 16 was used (due to the large size of the model, and memory constraints) during training. We used the standard Adam Optimizer [24], for training the model on a GPU grid for 20 epochs. The best model was selected based on the minimum loss in the validation set. The architecture was implemented using the PyTorch [32] library, and the training and evaluation components were implemented using the *Bob*² [1] library to make it easy to reproduce the results.

4. Experiments

Details of the databases used and the experimental results with the proposed approach are detailed in this section.

4.1. Databases

A wide variety of attacks are required to evaluate the performance of algorithms against unseen attacks. Most of the publicly available PAD datasets are limited to 2D print and replay attacks. Hence, we selected two publicly available datasets that contain a wide variety of 2D, 3D, and partial attacks, namely *HQ-WMCA* and *SiW-M* datasets.

***HQ-WMCA* dataset:** The High-Quality Wide Multi-Channel Attack (*HQ-WMCA*) dataset [20, 30] consists of 2904 short multi-modal video recordings of both bonafide and presentation attacks. The database includes both obfuscation and impersonation attacks and the attack categories present are print, replay, rigid mask, paper mask, flexible mask, mannequin, glasses, makeup, tattoo, and wig. The number of bonafide subjects available is 51 and the dataset contains several data streams captured synchronously such as color, depth, thermal, infrared (spectra), and short-wave infrared (spectra). In this work, we utilize only the RGB data stream from the dataset. The RGB videos are captured at a resolution of 1920×1200 .

We have created leave-one-out (LOO) attack protocols in *HQ-WMCA* by leaving out one attack type in the train and development set. The evaluation set consists of bonafide and the attack type which was left out. These sub-protocols constitute the zero-shot (or leave-one-out) protocols, which emulate the scenario of encountering an unseen attack type in a real-world scenario. We have also used the *grandtest* protocol which consists of attacks distributed in the train, development, and test sets (with disjoint identities across folds), specifically for cross-database performance evaluation.

²<https://www.idiap.ch/software/bob/>

SiW-M dataset: The Spoof in the Wild database with Multiple Attack Types (*SiW-M*) [27] again consists of a wide variety of attacks captured using an RGB camera. The number of subjects present is 493, with 660 *bonafide* and 968 attack samples with a total of 1628 files. There are 13 different sub-categories of attacks, collected in different sessions, pose, lighting, and expression (PIE) variations. The attacks consist of various types of masks, makeups, partial attacks, and 2D attacks. The RGB videos are available in 1920×1080 resolution³.

We use the leave-one-out (LOO) testing protocols available with the *SiW-M* [27] dataset for our experiments. The protocols available with the dataset consists of only *train* and *eval* sets. In each of the LOO protocols, the training set consists of 80% percentage of the bonafide data and 12 types of spoof samples. The test set consists of 20% of *bonafide* data and the attack which was left out in the training set. We created a subset of the train set (5%), as the *dev* set for model selection. In addition to the protocols available with the dataset, a *grandtest* protocol was also created (as done in [16]) specifically for cross-database testing with attacks distributed more or less equally across folds.

4.2. Metrics

For the evaluations in *HQ-WMCA* dataset, we have used the ISO/IEC 30107-3 metrics [22], Attack Presentation Classification Error Rate (APCER), and Bonafide Presentation Classification Error Rate (BPCER) along with the Average Classification Error Rate (ACER) in the *eval* set. We compute the threshold in the *dev* set for a BPCER value of 1%. The ACER in the *eval* set is calculated as the average of APCER and BPCER computed at this threshold.

For the *SiW-M* database, to enable comparison with other state-of-the-art methods, we follow the same method of reporting results as compared to [27]. Specifically, we apply a predefined threshold on the *eval* set of all the protocols. The ACER, APCER, and BPCER are computed using a fixed threshold of 0.5 on all the sub-protocols. Additionally, the equal error rate (EER) is also reported in the *eval* set.

For cross-database testing, Half Total Error Rate (HTER) is used following the convention in [16], which computes the average of False Rejection Rate (FRR) and the False Acceptance Rate (FAR).

HTER is computed in the *eval* set using the threshold computed in the *dev* set using the EER criterion.

4.3. Baseline methods

We have implemented three CNN-based baselines to compare with the proposed method. Since the proposed

³The *SiW-M* dataset is currently not publicly available due to a possible revision of the dataset. However, we have performed the experiments before the retraction of the dataset, and have obtained permission from the authors of the dataset to include the results in the current manuscript. The results correspond to the original release of *SiW-M* as used in [27].

method is based on transfer learning, we compare the proposed method with transfer learning from two popular architectures for image classification namely *ResNet* and *DenseNet* architectures. Besides, we have implemented *DeepPixBiS* architecture from literature which was specifically designed for presentation attack detection task. In addition to the implemented baselines, we compare the proposed approach with state-of-the-art methods from the literature in the *SiW-M* dataset. The details of the baseline methods implemented are given below.

ResNetPAD : Here we take the standard pre-trained *ResNet* model [19], specifically we used the *ResNet101* variant of the architecture. We replace the final layer with a new fully-connected layer making it suitable for binary classification. And while training, only the final fully connected layer is adapted.

DenseNetPAD: Similarly, here we take the standard *DenseNet* model [21] for fine-tuning. We used the *DenseNet161* variant of the architecture in our experiments. Here again, we replace the final layer with a new fully-connected layer for binary classification and during training, only the last layer is adapted.

DeepPixBiS: This is a CNN based system [14] which achieved good intra as well as cross-database performance in challenging OULU-NPU [7] dataset. The network was trained using both binary and pixel-wise binary loss functions. The usage of pixel-wise loss acts as an auxiliary loss function forcing the network to learn a robust classifier.

ViTranZFAS: This is our final proposed framework. Essentially, we take the pre-trained vision transformer model [11] and remove the final classification head. A new fully connected layer is added on top of the embedding followed by a sigmoid layer. The network is then trained using binary cross-entropy loss function, adapting only the final fully connected layer during training.

5. Experiments

We have conducted an extensive set of experiments in both intra and intra-dataset scenarios in both *HQ-WMCA* and *SiW-M* datasets. Specifically, we evaluate the baselines and the proposed approach in unseen attack environments (zero-shot) protocols as it indicates the performance of these PAD systems encountering real-world attacks that were not seen during training time.

Results in HQ-WMCA dataset: We have performed experiments using all the LOO protocols in the *HQ-WMCA* dataset and the results are tabulated in Table 1. The values reported are the ACER in the *eval* set corresponding to a threshold found from the *dev* set (using BPCER 1% criterion). It can be seen that the proposed achieves much better performance than the baseline methods, achieving an average ACER of $9.020 \pm 7.99\%$. This result is very promising

Table 1. Performance of the baseline systems and the proposed method in **unseen** protocols of *HQ-WMCA* dataset. The values reported are ACER’s obtained in the *eval* set with a threshold computed for BPCER 1% in *dev* set.

Method	Flexiblemask	Glasses	Makeup	Mannequin	Papermask	Rigidmask	Tattoo	Replay	Mean \pm Std
DenseNetPAD	28.20	45.50	36.60	0.40	6.90	12.70	4.60	32.40	20.91 \pm 15.76
ResNetPAD	39.10	42.00	41.20	2.80	0.50	21.30	28.60	21.50	24.62 \pm 15.33
DeepPixBiS [14]	5.80	49.30	23.80	0.00	0.00	25.90	13.60	6.00	15.55 \pm 15.76
<i>ViTranZFAS (FC)</i>	2.60	15.90	25.80	2.70	2.30	9.50	2.40	12.40	9.20\pm7.99

since only the last fully connected layer was retrained for classification.

Results in *SiW-M* dataset: The *SiW-M* dataset contains a wide variety of attacks. We have performed experiments with the zero-shot protocols and the results are tabulated in Table 2. In this database, the proposed approach achieves a large performance improvement, nearly half of the error rate as compared to the state-of-the-art methods. The proposed approach performs well on most of the sub-protocols and achieves a mean EER of 6.72 ± 5.66 %.

Analysis of training strategies: It was observed that fine-tuning the last layer alone was achieving state-of-the-art performance. Here we examine the effectiveness of retraining different sets of layers in the *HQ-WMCA* dataset. We have considered three different settings for this study they are,

- *ViTranZFAS (FC)*: Here, only the last fully connected layer is retrained, all the other layers are frozen, this corresponds to the fine-tuning scenario from a pre-trained model.
- *ViTranZFAS (ALL)* layers: All the layers from the architecture are adapted during training.
- *ViTranZFAS (E+FC)*: Here both the first embedding layers as well as the final fully connected layer are adapted.

The results for this set of experiments are shown in Table 3. Clearly, adapting only the FC layer achieves the best results. Given the large amount of data used for training the pre-trained model, adapting other layers appears to be prohibitive with a limited amount of training data.

Figure 2 shows the t-SNE plots from the Vision Transformer embeddings (from the pre-trained models) on both *HQ-WMCA* and *SiW-M* datasets individually and together (for the *eval* set in the corresponding *grandtest* protocols). It can be seen that there is already a good amount of separability between bonafide and the spoof samples in the feature space. This could explain the good performance of the proposed method just by adapting the final fully connected layer.

Visualization of different classes: To further understand the features contributing to the decisions, we computed the relevancy maps for different classes. In [9], the authors

proposed a way to visualize the relevancy maps for Transformer networks. Essentially, their method assigns local relevance based on the deep Taylor decomposition which propagates the relevancy scores through the layers. This method achieved the state of the art results compared to other methods, for computing the relevancy maps for Transformer networks. We have computed the relevancy maps for the fine-tuned Vision Transformer model (*ViTranZFAS (FC)*). The relevance maps for various attack types are shown in Fig. 3. Interestingly, in “Makeup” and “Tattoo” attacks, the network correctly identifies the regions of importance. For attacks like “Replay” and “Rigid mask” and the network struggles as the discriminative region is not localized, the spoof traces a spread out throughout the face in such attacks.

Cross-database evaluations : From the previous sections, it was clear that the proposed approach achieves much better performance as compared to the state-of-the-art methods in unseen attack scenarios. One main issue of PAD methods has been poor cross-database generalization, which is essential to ensure reliable performance in real-world deployment scenarios. To evaluate the generalization, we performed cross-database evaluations between *HQ-WMCA* and *SiW-M* datasets. For each dataset, we trained the model using the *grandtest* protocol of the corresponding dataset, and the resulting model is evaluated using the *grandtest* protocol of the other dataset. The results are tabulated in table 4. From the results, it can be seen that the proposed approach improves the cross-database performance by a large margin indicating the generalizability of the proposed approach.

Computational Complexity: Here we compare the parameters, and complexity of the baseline and the vision transformer model. The comparison is shown in Table 5. It can be seen that the vision transformer model requires more computational and parameters as compared to the baselines. Though the network is complex, we just retrain just the last fully connected layer with just 768 neurons in our transfer learning setting making the training far easier. Distillation of the model [35] to reduce the complexity could be a possible direction to address this limitation.

Discussions: From the experimental results in both *HQ-WMCA* and *SiW-M* datasets, it can be seen that the proposed method achieves state-of-the-art performance in chal-

Table 2. Performance of the proposed framework in the leave one out protocols in *SiW-M* dataset.

Methods	Metrics (%)	Replay	Print	Mask Attacks					Makeup Attacks			Partial Attacks			Average
				Half	Silicone	Trans.	Paper	Manne.	Obfusc.	Imperson.	Cosmetic	Funny Eye	Paper Glasses	Partial Paper	
Auxiliary [26]	APCER	23.7	7.3	27.7	18.2	97.8	8.3	16.2	100.0	18.0	16.3	91.8	72.2	0.4	38.3 ± 37.4
	BPCER	10.1	6.5	10.9	11.6	6.2	7.8	9.3	11.6	9.3	7.1	6.2	8.8	10.3	
	ACER	16.8	6.9	19.3	14.9	52.1	8.0	12.8	55.8	13.7	11.7	49.0	40.5	5.3	
	EER	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	9.4	21.4	18.6	4.0	
Deep Tree Network [27]	APCER	1.0	0.0	0.7	24.5	58.6	0.5	3.8	73.2	13.2	12.4	17.0	17.0	0.2	17.1 ± 23.3
	BPCER	18.6	11.9	29.3	12.8	13.4	8.5	23.0	11.5	9.6	16.0	21.5	22.6	16.8	
	ACER	9.8	6.0	15.0	18.7	36.0	4.5	7.7	48.1	11.4	14.2	19.3	19.8	8.5	
	EER	10.0	2.1	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	19.8	20.5	8.8	
MCCNN (BCE+OCCL)-GMM [16]	APCER	11.79	9.53	3.12	3.70	39.20	0.00	3.12	44.57	0.00	21.60	19.34	35.55	0.00	14.7 ± 15.9
	BPCER	13.44	16.15	16.26	20.23	11.11	13.74	8.66	15.23	12.67	10.42	14.31	18.40	27.33	
	ACER	12.61	12.84	9.69	11.97	25.16	6.87	5.89	29.90	6.34	16.01	16.83	26.97	13.66	
	EER	12.82	12.94	11.33	13.70	13.47	0.56	5.60	22.17	0.59	15.14	14.40	23.93	9.82	
DenseNetPAD	APCER	28.32	11.00	26.81	23.84	39.88	0.04	3.22	70.69	0.01	38.21	72.66	53.37	3.46	28.58 ± 24.48
	BPCER	12.48	12.44	12.63	13.63	12.31	13.18	13.81	11.50	13.93	11.60	12.59	12.00	12.48	
	ACER	20.40	11.72	19.72	18.74	26.10	6.61	8.51	41.10	6.97	24.90	42.63	32.69	7.97	
	EER	16.49	12.10	15.43	17.03	19.91	5.58	10.44	24.21	3.67	17.66	29.12	22.26	9.78	
ResNetPAD	APCER	32.08	18.88	34.64	21.22	33.96	0.00	2.78	94.38	0.00	35.38	72.92	18.83	2.72	28.29 ± 27.22
	BPCER	9.47	9.48	10.51	10.95	10.50	11.33	10.79	9.39	11.03	10.17	10.79	11.09	10.68	
	ACER	20.78	14.18	22.57	16.09	22.23	5.66	6.79	51.88	5.51	22.77	41.86	14.96	6.70	
	EER	15.17	11.81	18.25	14.69	16.87	1.96	7.95	33.27	4.19	17.72	28.85	12.86	7.86	
DeepPixBiS [14]	APCER	19.18	8.97	1.74	21.30	60.68	0.00	1.00	100.00	0.00	26.90	64.66	77.52	0.29	29.4 ± 34.4
	BPCER	8.70	7.63	11.03	11.76	10.27	8.85	8.63	10.53	11.60	10.99	10.31	10.23	7.10	
	ACER	13.94	8.30	6.38	16.53	35.47	4.43	4.81	55.27	5.80	18.95	37.48	43.87	3.69	
	EER	11.68	7.94	7.22	15.04	21.30	3.78	4.52	26.49	1.23	14.89	23.28	18.90	4.82	
<i>ViTranZFAS (FC)</i>	APCER	38.27	5.81	5.00	4.62	5.47	0.00	0.32	12.55	0.00	18.32	61.81	0.29	0.13	11.74 ± 17.75
	BPCER	4.82	5.87	6.27	5.52	6.33	5.68	6.02	6.22	6.66	5.62	5.46	7.03	5.50	
	ACER	21.55	5.84	5.63	5.07	5.90	2.84	3.17	9.38	3.33	11.97	33.63	3.66	2.82	
	EER	15.20	5.84	5.80	4.99	5.95	0.12	3.25	9.89	0.46	10.76	20.19	2.96	1.97	

Table 3. Performance of the Vision Transformer network when fine tuning different set of layers in **unseen** protocols of *HQ-WMCA* dataset. The values reported are ACER's obtained in *eval* set with a threshold computed for BPCER 1% in *dev* set.

Adapted Layers	Flexiblemask	Glasses	Makeup	Mannequin	Papermask	Rigidmask	Tattoo	Replay	Mean ± Std
<i>ViTranZFAS (ALL)</i>	2.40	44.90	21.10	0.00	0.20	21.60	0.60	25.90	14.59 ± 15.42
<i>ViTranZFAS (E+FC)</i>	5.50	47.00	23.40	1.90	19.40	11.60	2.10	11.50	15.30 ± 13.99
<i>ViTranZFAS (FC)</i>	2.60	15.90	25.80	2.70	2.30	9.50	2.40	12.40	9.20 ± 7.99

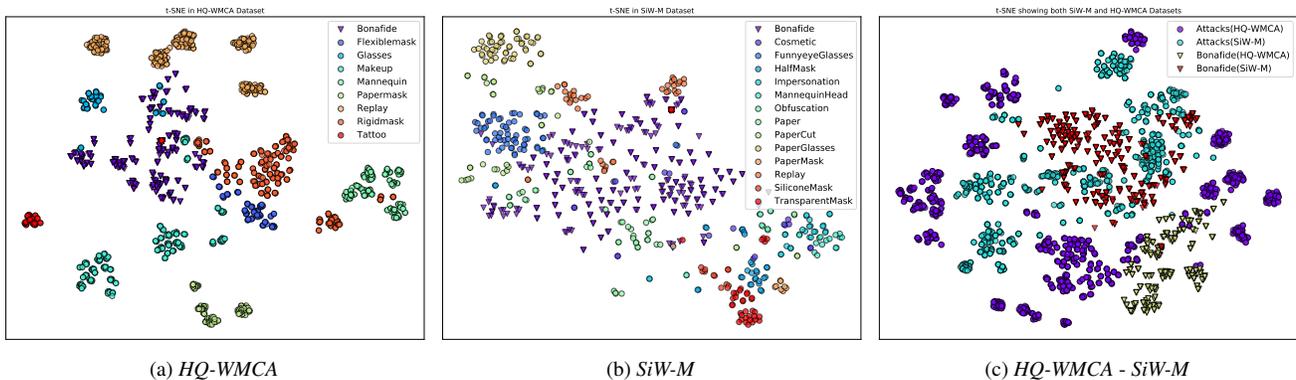


Figure 2. t-SNE plots corresponding to the 768-dimensional feature from the Vision Transformer, the plot already shows some separability between bonafide and attacks in the features extracted from the pre-trained model.

lenging unseen attack scenarios. Surprisingly the proposed method achieves excellent cross-database generalization as well. Typical PAD models have a tendency to overfit to the nuances in specific datasets rather than focusing on the reliable spoof cues, resulting in poor generalization in cross-database scenarios. Using the pre-trained model provides a good prior, and training strategy adapting a minimal subset of layers reduces the chances of overfitting, achieving good performance in the challenging scenarios. The inherent properties of the vision transformers also boost the per-

formance. Self-attention as opposed to convolutions helps to attend to all pairwise interactions in the lower layers itself. The large datasets used for pre-training the vision transformer models also improve the robustness. As shown in Fig. 3, the model correctly focuses on the discriminative features.

6. Conclusions

In this work, we have shown the effectiveness of the vision transformer network for the zero-shot face anti-

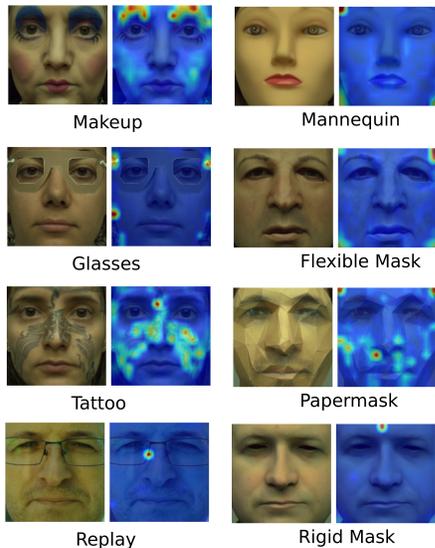


Figure 3. Relevancy maps for different types of attacks with *ViTranZFAS (FC)* model in the grandtest protocol in *HQ-WMCA* dataset.

Table 4. The results from the cross-database testing between *SiW-M* and *HQ-WMCA* datasets. HTER (%) values computed in *eval* set for threshold computed in *dev* set using EER criteria are reported in the table.

Method	trained on <i>HQ-WMCA</i>		trained on <i>SiW-M</i>	
	tested on <i>HQ-WMCA</i>	tested on <i>SiW-M</i>	tested on <i>SiW-M</i>	tested on <i>HQ-WMCA</i>
DenseNetPAD	11.40	27.5	10.4	29.3
ResNetPAD	13.50	25.7	10.4	29.4
DeepPixBiS [14]	4.60	25.6	14.7	38.1
<i>ViTranZFAS (FC)</i>	5.60	14.7	6.0	12.7

Table 5. Computational and parameter complexity comparison

Model	Compute	Parameters
ResNetPAD	7.85 GMac	42.5 M
DenseNetPAD	7.82 GMac	26.47 M
MCCNN(RGB) [16]	10.88 GMac	37.73 M
DeepPixBiS [14]	4.64 GMac	3.2M
<i>ViTranZFAS</i>	16.85 GMac	85.8 M

spoofing task. Essentially, just fine-tuning a pre-trained vision transformer model for the PAD task was sufficient to achieve the state-of-the-art performance in *HQ-WMCA* and *SiW-M* datasets. In addition to excellent performance in unseen attacks, the proposed approach outperforms the state-of-the-art methods in cross-datasets evaluations by a large margin, indicating the efficacy of the proposed approach in generalizing to both unseen attacks and domains. The vision transformers could prove to be very beneficial in dealing with the current limitations of presentation attack detection systems. The datasets, source code, and protocols used are made available publicly to enable the further extension of the work.

To summarize, in this work we show that merely fine-tuning the last fully connected layer in vision transformers achieves state-of-the-art performance in both unseen attack and cross-database scenarios. Extensive evaluations show the effectiveness of the method. The superior performance in addressing two of the challenging issues (unseen attack and cross-database generalization) in the PAD task with minimal fine-tuning holds the potential to address the issues with PAD models. We hope that this work will motivate the biometrics community to investigate transformer models further.

Acknowledgment

Part of this research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017-17020200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

- [1] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel. Continuously reproducing toolchains in pattern recognition and machine learning experiments. In *International Conference on Machine Learning (ICML)*, Aug. 2017.
- [2] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *Biometrics (IJCB), 2011 international joint conference on*, pages 1–7. IEEE, 2011.
- [3] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. Face anti-spoofing using patch and depth-based cnns. In *Biometrics (IJCB), 2017 IEEE International Joint Conference on*, pages 319–328. IEEE, 2017.
- [4] D. Bahdanau, K. H. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.
- [5] S. Bhattacharjee, A. Mohammadi, and S. Marcel. Spoofing deep face recognition with custom silicone masks. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018.
- [6] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing based on color texture analysis. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 2636–2640. IEEE, 2015.
- [7] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *Automatic Face & Gesture Recog-*

- tion (*FG 2017*), *2017 12th IEEE International Conference on*, pages 612–618. IEEE, 2017.
- [8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- [9] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. *arXiv preprint arXiv:2012.09838*, 2020.
- [10] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Proceedings of the 11th International Conference of the Biometrics Special Interest Group*, number EPFL-CONF-192369, 2012.
- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
- [12] N. Erdogmus and S. Marcel. Spoofing face recognition with 3d masks. *IEEE transactions on information forensics and security*, 9(7):1084–1097, 2014.
- [13] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE transactions on image processing*, 23(2):710–724, 2014.
- [14] A. George and S. Marcel. Deep Pixel-wise Binary Supervision for Face Presentation Attack Detection. In *International Conference on Biometrics (ICB)*, 2019.
- [15] A. George and S. Marcel. Can your face detector do anti-spoofing? face presentation attack detection with a multi-channel face detector, 6 2020.
- [16] A. George and S. Marcel. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, pages 1–1, 2020.
- [17] A. George and S. Marcel. Cross modal focal loss for RGBD face anti-spoofing. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [18] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, pages 1–1, 2019.
- [19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [20] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, and S. Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 2020.
- [21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [22] Information technology –International Organization for Standardization. Standard, International Organization for Standardization, Feb. 2016.
- [23] A. K. Jain and S. Z. Li. *Handbook of face recognition*. Springer, 2011.
- [24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.
- [25] R. Liu, Z. Yuan, T. Liu, and Z. Xiong. End-to-end lane shape prediction with transformers, 2020.
- [26] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018.
- [27] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu. Deep tree learning for zero-shot face anti-spoofing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [28] J. Määttä, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *Biometrics (IJCB), 2011 international joint conference on*, pages 1–7. IEEE, 2011.
- [29] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans. *Handbook of biometric anti-spoofing : Presentation attack detection*. Editors: Marcel, S., Nixon, M.S., Fierrez, J., Evans, N. (Eds.); Springer International Publishing, 2018, 2nd ed.; ISBN: 978-3319926261, 09 2018.
- [30] Z. Mostaani, A. George, G. Heusch, D. Geissenbuhler, and S. Marcel. The high-quality wide multi-channel attack (hq-wmca) database, 9 2020.
- [31] O. Nikisins, A. George, and S. Marcel. Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.
- [32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
- [33] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch. Extended multispectral face presentation attack detection: An approach based on fusing information from individual spectral bands. In *Information Fusion (Fusion), 2017 20th International Conference on*, pages 1–6. IEEE, 2017.
- [34] H. Steiner, A. Kolb, and N. Jung. Reliable face anti-spoofing using multispectral swir imaging. In *Biometrics (ICB), 2016 International Conference on*, pages 1–8. IEEE, 2016.
- [35] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.
- [36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [37] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.
- [38] R. Wightman. PyTorch Image Models. <https://github.com/rwightman/pytorch-image-models>, 2020.
- [39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.