IDIAP RESEARCH REPORT

ASSESSING THE RELIABILITY OF BIOMETRIC AUTHENTICATION ON VIRTUAL REALITY DEVICES

Ketan Kotwal Gokhan Ozbulak Sébastien Marcel

Idiap-RR-04-2024

JULY 2024

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Assessing the Reliability of Biometric Authentication on Virtual Reality Devices

Ketan Kotwal[1], Gökhan Özbulak[1], *and* Sébastien Marcel[1,2]
[1] Idiap Research Institute, Switzerland
[2] University of Lausanne, Switzerland
{ketan.kotwal, gokhan.ozbulak, sebastien.marcel}@idiap.ch

Abstract

Recent developments in Virtual Reality (VR) headsets have unlocked a plethora of innovative use-cases, many of which were previously unimaginable. However, as these use-cases, such as personalized immersive experiences, necessitate user authentication, ensuring robustness and resistance to spoofing attacks becomes imperative. The absence of appropriate dataset has constrained our understanding and assessment of VR devices' vulnerability to presentation attacks. To address this research gap, we introduce a new periocular video dataset acquired from a VR headset (Meta Quest Pro), comprising 900 genuine and 996 presentation attack videos, each spanning 10 seconds. The bona-fide videos consist of variations in terms of gaze and glasses; while the attacks are constructed with 6 different types of instruments. Additionally, we evaluate the performance of two prominent CNN architectures trained using various configurations for detecting presentation attacks in the newly created dataset, VRPAD. Our benchmarking on VRPAD reveals the presence of spoofing threats in VR headsets. While baseline models exhibit considerable efficacy in attack detection, substantial scope exists for improvement in detecting attacks on periocular videos. Our dataset will be a useful resource for researchers aiming to enhance the security and reliability of VR-based authentication systems.

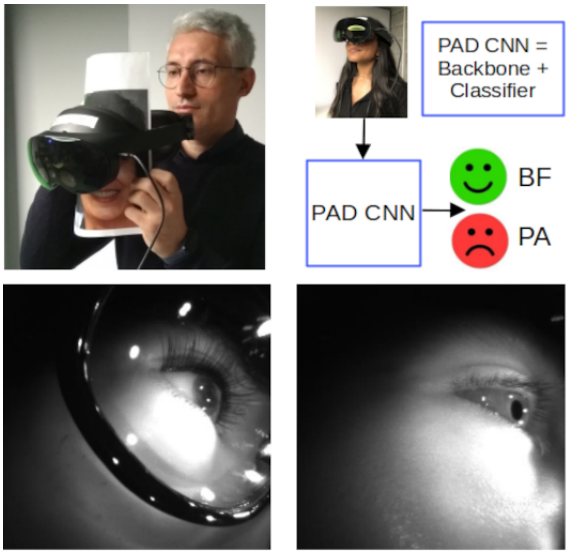


Figure 1. In top left, a user attempting to spoof the Iris/Periocular authentication of the Virtual Reality (VR) headset by presenting a print-out of face of another real or synthetic identity. The cameras, mounted on the temple region of the device, are intended to capture the periocular regions of the wearer of headset. This picture is representational as controlled exposure of incident light is necessary to record good quality attacks. The bottom row depicts attack (left) and *bona-fide* (right) samples from the VRPAD dataset.

1. Introduction

In recent years, the advancement in the wearable devices such as Head-Mounted Display (HMD) made it possible to interact with real and/or virtual environments for specific purposes. The HMD device operated in such virtual on real (Augmented Reality), virtual only (Virtual Reality) and virtual/real blended (Mixed Reality) environments enable the people to play computer games, interact with multimedia including video and train for profession-purpose. There are many vendors such as Meta, Apple, Microsoft and HTC focusing on the development of these kind of devices in different real/virtual level.

Beyond gaming, the HMD devices have many application areas ranging from healthcare to communication. In healthcare, it can be adopted for use in surgical training and planning [12, 22], and in more specifically for the ophthalmology and psychology, it may be possible to diagnose the eye diseases [24], as well as to analyze the human emotion/behavior [3] based on the data collected by such devices. In communication, the avatar as digitization of the real people in the Metaverse, which is defined as the virtual world's access by HMD devices, provides a way of interaction for people to communicate with each other on virtual platforms [4]. This kind of communication requires the assurance regarding the validity on person on the other side,

and needs additional security layers such as identity recognition using biometric traits. The use of iris or periocular region, being inherently visible to the HMD, as biometric trait can be natural choice in this regard.

The existence of the biometric systems is vital for these devices in order to provide secure access to the personal information regarding to healthcare or privacy. The introduction of such systems also requires the robustness against presentation attacks that are realized over various instruments (mask, paper etc.). A Presentation Attack Detection (PAD) system must be co-exist with the biometric system in order to keep unauthorized access, which mimics as an authorized person over printed paper etc., away. Although it's an achievable data-driven machine learning task because of the plenty number of the PAD dataset in the mainstream biometric recognition [6, 18], to the best knowledge of us, there is no PAD dataset collected specifically from HMD devices.

With emergence of VR devices, and possible diverse usage, it is important to envisage upcoming challenges and conduct research to address these well in advance. The critical missing factor is no such data is publicly available given the novelty or recentness of such data. In this work, we address both challenges associated: first, we create a dataset of nearly 1,900 periocular videos, nearly 10 s long, captured by the headset cameras of Meta Quest Pro (a VR headset released by Meta, formerly Facebook). This dataset consists of recordings of 25 *bona-fide* users with still gaze, moving gaze, as well as partially closed eyes. For selected recordings, we construct a presentation attack (PA) by fake 3D eyeballs, print-outs of periocular regions of other enrolled users as well as that of synthetic identities. This dataset-**V**irtual **R**eality **P**resentation **A**ttack **D**ataset (**VRPAD**)- will be publicly released[1] to the research community for furthering work in this area. Secondly, we conduct PAD assessment of the newly created VRPAD using two most commonly used convolutional neural network (CNN) architectures: ResNet and MobileNet. For each architecture, we select three different strategies based on which components of the CNN are adapted (trained/ finetuned) or regarded as constant (frozen). For each of these combinations, we evaluate their robustness towards detection of PA from the VRPAD.

The main contributions of the paper can be summarized as below:

- We create a dataset of 900 iris/ periocular videos of 25 subjects captured by the cameras of a VR headset, specifically the Meta Quest Pro. We also capture nearly 1,000 videos of presentation attacks constructed using fake eyeballs, printouts, and synthetic eyes using a variety of masks and mannequins.
- We conduct PAD assessment of the newly created dataset using two state-of-the-art architectures, by training them in different strategies. Results of our experiments indicate that while the existing CNN architectures, along with pretrained weights, are able to detect the PAs to a large extent, there is a systematic opportunity for improvement.

[1]URL will be shared upon acceptance of the paper.

2. Related Work

In this section, we first provide an overview of recent advances in iris/ periocular PAD methods, then we discuss in brief some existing datasets acquired using HMD devices.

Iris/Periocular PAD: The PAD for iris or periocular regions as biometric trait is a well-studied area of research. In [19], Sharma *et al.* proposed a DenseNet based PAD system in order to expose the complicated features of the iris stroma by the multi-resolution analysis capability of the DenseNet. Similar to [19], Swarup *et al.* also used the DenseNet combined with attention mechanism between the dense blocks of the DenseNet [20]. Li *et al.* leveraged the frequency domain to alleviate the cross-dataset performance degradation because of the limited data on the *bona-fide* samples [14]. Fang *et al.* proposed an attention-based deep pixel-wise method, which detects the regions effecting the performance of the PAD, in [9] for infrared and visible domains. In [1], the problem of iris PAD is considered for the contact lens scenario and a generalized CNN topology is used combined with early and late fusion strategies in order to have robust PAD methods in cross-dataset setting. Dhar *et al.* proposed a multitask system for both eye authentication and PAD [7]. They mitigated the problem of forgetting in the Multitask Learning (MTL) by introducing the distillation-based approach. In [23], a Generative Adversarial Network (GAN) based approach is used in order to generate synthetic PA samples that are underrepresented in the training set. Agarwal *et al.* proposed a PAD system by combining gener-covariate based classifiers in [2]. In the same study, fairness based observations are also shared related with gender bias in the PAD datasets. In [15], Li *et al.* proposed a PAD system for better generalization in a cross-dataset setting by exposing domain-invariant and domain-specific features in the samples.

HMD Datasets: HMD based datasets are severe in the literature as the topic is relatively new. The most recent study is proposed as the OpenEDS dataset by Garbin *et al.* [10]. This dataset consists of $356k$ infrared images with an image size of 640×400 collected from 152 subjects. The another recent dataset is the NVGaze proposed by Kim *et al.* [11] and it contains 2.5 million infrared images with dimensions of 640×480 for 30 subjects. Tonsen *et al.* proposed a similar dataset named as the LPW [21] with 22 subjects in a head-mounted fashion and it consists of $130k$ images with a spatial resolution of 640×480. The Point of Gaze dataset, by McMurrough *et al.* [16], is collected from

20 subjects and contains images with a spatial resolution of 768×480 pixels.

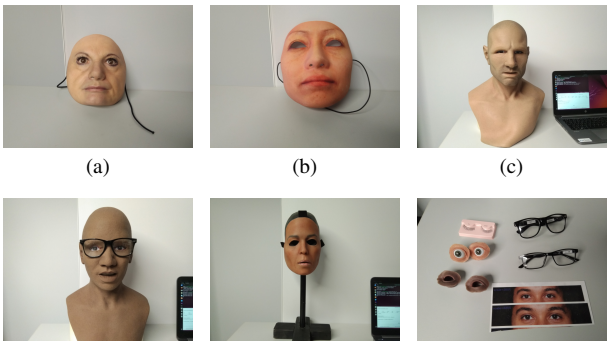
3. New Dataset: VRPAD

This section describes the details regarding to the VRPAD dataset including the statistics for *bona-fide* and attack samples, the PA instruments used during the data collection session and the protocol information.

3.1. Description

The VRPAD dataset, to the best of our knowledge, is the first periocular PAD dataset acquired from HMD devices such as Meta Quest Pro. Since the environment of data acquisition is highly controlled (the HMD closely fits around the head of the subject), we recorded every identity, either *bona-fide* or PAIs, in one session. Each recorded sample is composed of two sub-samples referring to Near Infrared (NIR) cameras of Meta Quest Pro looking into left and right eyes (temples) of the identity.

During *bona-fide* session, each participant was briefly informed about the project, and was required to sign a consent form. As seen in Fig. 3, each subject was recorded in two sub-sessions: one with wearing glasses and another without. For each sub-session, the subject was asked to maintain three specific conditions as *still gaze*, *moving gaze*, and *partially closed* eyes. We recorded each video for nearly $10s$ at 72 *fps*. Post recording, we often observed that during initial first second or so, several recordings were over-exposed, and thus, initial 70 frames were discarded. The videos, thus, have around 650 frames and spatial resolution is 400×400.



(a) (b) (c) (d) (e) (f)

Figure 2. PA instruments for VRPAD: (a) Rigid masks with own eyes, (b) rigid masks with fake eyeballs, (c) flex masks with print attacks, (d) flex masks with print attacks, (e) flex masks with fake eyeballs, and (f) auxiliary instruments (fake eyeballs, prints with synthetic eyes, eyelashes, glasses).

We constructed presentation attacks with different combinations of Presentation Attack Instruments (PAIs) to create a wide range of attack dataset. The summary of combination of 3D masks (and mannequins) and fake eyes considered to create attacks is provided in Table 1. Fig. 2 depicts some samples of PAIs which include rigid masks with own eyes, rigid masks with fake 3D eyeballs, generic flex masks with printouts of synthetic eyes, custom flex masks with fake 3D eyeballs, and attacks constructed from prints of *bona-fide* samples. Some auxiliary instruments such as fake eyeballs, eyelashes and glasses were also used to bring more variations and realism to PAIs. Some samples from PA sessions can be examined in Fig. 4.

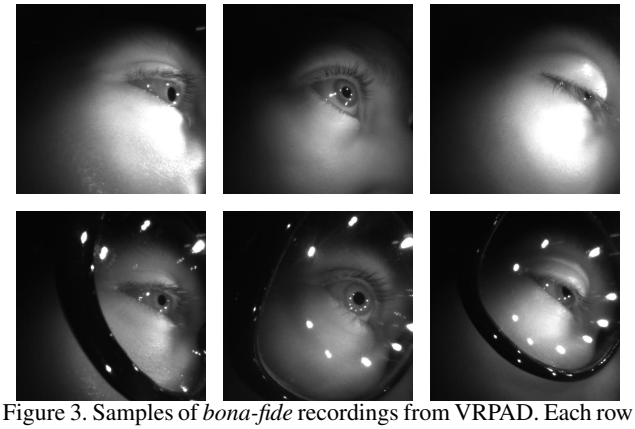
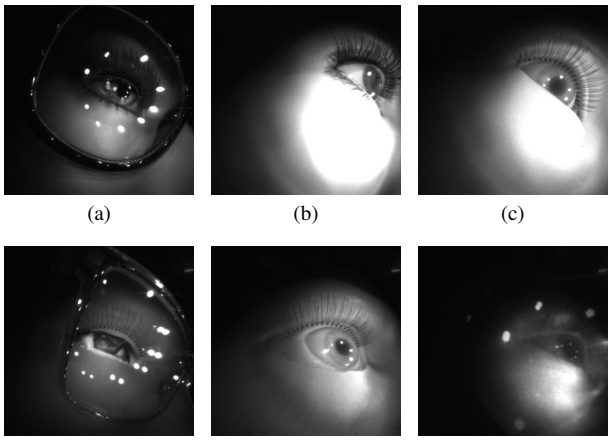


Figure 3. Samples of *bona-fide* recordings from VRPAD. Each row presents a sample of still gaze, moving gaze, and partially closed eyes (from left to right). Top and bottom rows refer to recordings without and with glasses, respectively.

3.2. Dataset Statistics

Our dataset comprises of video recordings of 25 *bona-fide* subjects. For each subject, a total of 36 video samples were captured: which are combinations of 3 gaze scenarios, 2 glass/no-glass, and 3 repetitions from each left and right camera. In total, 900 *bona-fide* videos were recorded. Subsequently, for each subject, we chose a suitable near-



(a) (b) (c) (d) (e) (f)

Figure 4. Samples of PA recordings from VRPAD. (a) Mannequins with own eyes, (b) rigid masks with own eyes, (c) rigid masks with fake eyeballs, (d) flex masks with print attacks, (e) flex masks with fake eyeballs, and (f) print attacks from *bona-fide* samples.

frontal frame from with and without glass recordings. A print-out of this frame from a laser printer (which is visible in NIR spectrum) was used to create print-attacks. For each eye (left and right), we recorded three repetitions, with and without glasses for such attacks; thereby obtaining a set of 300 attacks videos that can be assessed for not only *obfuscation*, but also for vulnerability of recognition system.

For PAs, each instrument was recorded 3 times with and without glasses. Thus, for each fake identity, we obtained 18 attack videos. Our VRPAD encompasses 7 unique mannequins that were used to obtain ($7 \times 3 \times 2 =$) 86 videos. Similarly, for two types of custom rigid masks (ones with own eyes and ones fake 3D eyeballs), we obtained 120 and 60 attack presentations. For flexible masks with printed eyes and fake 3D eyeballs, we recorded 240 and 192 videos, respectively.

3.3. Experimental Protocol

To conduct PAD assessment and establish baselines on the VRPAD, we created an experimental protocol. We divided the VRPAD into `train`, `val`, and `test` partitions. We ensured that these partitions are disjoint in terms of identities, and nearly equal in volume (*i.e.*, 33% of total dataset). With uniform sampling, we selected every 10-th frame of each video for processing. It should be noted that all experiments described in this work have been conducted at frame-level. That is, each constituent frame is regarded as an independent data, irrespective of the implicit correlation frames of same video may have. Detection of attacks from video data, by exploiting temporal information, is not considered in this work. For frame-level PAD assessment, the `train` partition of VRPAD consists of 37,280 frames from 588 videos. The `val` partition has 38,563 frames from 512 videos—which have different identities as that from the training partition. For the `test` partition, our experimental protocol consists of 43,648 frames from 696 videos (*bona-fide* and attacks combined). Some samples which were not recorded correctly, due to technical or manual glitches were removed after manual inspection[2].

Fig. 3 shows samples of *bona-fide* recordings, and Fig. 4 presents different types of PAs.

4. PAD Assessment of VRPAD

Initially, we provide the specifics of the experimental setup, models, and evaluation metrics used. Subsequently, we discuss the outcomes of the PAD experiments conducted on the VRPAD. These experiments are combinations involving two backbone architectures and three different models (training configurations).

[2]We plan to re-record some videos that were removed in the present version. These videos are less than 3% of overall dataset, and similarly distributed across different partitions

4.1. Experimental Setup

Since the periocular region is considered as a biometric trait to be assessed for PAD, we adopt the entire frame as the input. From samples in Figs. 3 and 4, it can be observed that the frame of Meta Quest Pro encompasses the eyes and surrounding areas, validating the use of the entire frame as the periocular region. Additionally, due to the variability in eye positions (relative to the frame borders), cropping a fixed region cannot guarantee a more precise and confined periocular region. We conducted experiments to identify iris regions to dynamically obtain a smaller periocular region. However, our preliminary experiments indicated that commonly available eye detectors, such as OpenCV, are not reliable for accurate detection of eyes from VRPAD samples. Although we did not conduct a detailed analysis of detection failures, potential reasons include:

- **Non-frontal Views:** Most state-of-the-art eye or iris detectors are trained to process frontal views. The non-frontal, oblique nature of the VRPAD samples, resulting in elliptical iris shapes instead of circular ones, leads to poor detections.
- **Domain Shift:** Common models for eye detection are trained on RGB or visible spectrum data, whereas the VRPAD samples are acquired in the NIR spectrum, causing a domain shift.
- **Size:** The samples in VRPAD are relatively small, with an average iris diameter of less than 60 pixels.

Nevertheless, qualitatively and quantitatively improved iris detection from single or multiple frames remains an open research problem. We anticipate that advancements in iris detection will yield positive impacts on PAD or any pertinent applications of the VRPAD. We did not perform specific studies, such as domain adaptation or transfer learning, on eye/iris detection as the focus of this work is confined to understanding the threat of PAs to VR-based authentication systems.

PAD CNN Models: The PAD CNN models can be considered as a sequential combination of a backbone and a classifier. The backbone comprises a convolutional stem derived from a well-established CNN architecture, while the classifier is a conventional feed-forward neural network (FF-NN) tailored for binary classification. Prior to being fed into the classifier, the output of the backbone may undergo processing steps such as flattening, batch normalization, or dimensionality reduction. The output of the classifier is a scalar obtained via a Sigmoidal activation function, constraining its range to [0, 1]. This output indicates the probability of a sample being *bona-fide*. In this work, we have investigated two widely recognized architectures as backbones for PAD CNN:

Type	# Identities	# Videos	Subtype	Attack Types
bona-fide	25	900	[still gaze, moving gaze, half closed] $\times$ [glass, no glass]	
Presentation Attacks	7	84	Mannequins [2]	Own eyes (same material)
	10	120	Custom rigid mask with own eyes [3]	Own eyes (same material)
	5	60	Custom rigid mask [4]	Fake 3D eyeballs
	20	240	Generic flexible masks [5]	Print attacks (synthetic data)
	16	192	Custom silicone masks [6]	Fake 3D eyeballs
Vulnerability Attacks	25	300	Print attacks [7]	Prints of real data

Table 1. Details of *bona-fide* and different types of PAs from VRPAD. Each video is recorded at 72 FPS for approximately 10s. The numbers in [ ] brackets are used to refer type of PAI (called attack series, hereafter) in rest of the discussions.

- **ResNet-34:** This architecture represents a 34-layer variant of the improved ResNet architecture [8]. The architectures based on residual connections have demonstrated superior performance across various applications. They facilitate the learning of subtle data features, such as micro-textures potentially containing discriminative information related to PAs, while mitigating issues like vanishing gradients.
- **MobileFaceNet:** This architecture belongs to a family of compact architectures specifically optimized for operation in resource-constrained environments in handheld devices. This characteristic makes MobileNets an attractive choice for scenarios where the PAD model may operate on an HMD device. With use of depthwise separable convolutions, MobileNets achieve high performance while maintaining a low parameter count. In this work, we utilize MobileFaceNet [5], a variant of MobileNets tailored for face recognition tasks, as our backbone.

Training Configurations: Based on the selection of layers within the PAD CNN and the chosen training strategy, we conducted experiments with three distinct configurations outlined below:

- **Model1:** The entire PAD CNN (*i.e.* both the backbone and classifier) is trained from scratch on the `train` partition of the VRPAD. The models are initialized with random weights.
- **Model2:** The backbone utilizes weights from a pretrained model and remains fixed (frozen) throughout training, while the classifier is trained from scratch. In this configuration, the backbone serves as a fixed feature extractor, providing an *embedding* of the input sample to the classifier for detection of PAs.
- **Model3:** Similar to the previous configuration, the backbone employs pretrained weights while the classifier is initialized with random weights. However, during training, the entire PAD CNN (including the backbone) undergoes training or fine-tuning for the PAD objective. No layers of the backbone are frozen.

We selected these configurations to provide an overview of commonly employed strategies for training PAD models. In the case of pretrained weights (Model2 and Model3), we opted for models pretrained for face recognition for two reasons[3]. Firstly, since the periocular region is a subset of the face, the initial layers of the backbone may be more effective at learning the desired features of the present use-case. Secondly, the efficacy of face recognition CNNs toward PAD via transfer learning has been extensively studied [13, 17].

During training, frames from various videos (*bona-fide* + PA) are randomly grouped to form training batches of 64 samples. Preprocessing and data augmentation involved resizing the samples to 112×112, as required by the PAD CNN backbones. The output of the backbone, an intermittent layer pf the PAD CNN, was configured to be 512-d. We employed random horizontal flipping and random rotation up to 20° to improve generalization. Additionally, dropout was applied across multiple layers of the PAD CNNs.

An SGD-based optimizer with an initial learning rate of $1e$-3 was utilized alongside a step-wise rate scheduler. The weight decay and momentum of the optimizer were set to $1e$-4 and 0.90, respectively.

PAD Pipeline:[4] To establish PAD baselines, we have used to the conventional pipeline where the preprocessed image (or frame) of the input presentation is fed into a CNN, which yields a PAD score. This score, within a predefined

[3]ResNet: https://github.com/deepinsight/insightface, MobileFaceNet: https://github.com/yeyupiaoling/Pytorch-MobileFaceNet

[4]The source code for all PAD experiments described in this work will be released publicly after acceptance of the paper.

range, is then thresholded to obtain a binary decision (*bona-fide* or attack). Our training utilizes the `train` partition, while the `val` partition is employed for model validation at each training epoch and to determine the score threshold τ. Finally, the results are reported on the `test` partition of the VRPAD.

Evaluation Metrics: We report the performance of aforementioned configurations toward PAD on the VRPAD using the following evaluation metrics:

- **APCER** (Attack Presentation Classification Error Rate): APCER measures the proportion of PAs incorrectly classified as genuine (*bona-fide*). For N_{PAIs} PAIs, APCER is calculated as:

$$\text{APCER}_{\text{PAIs}} = 1 - \frac{1}{N_{\text{PAIs}}} \sum_{i=1}^{N_{\text{PAIs}}} \text{score}_i, \qquad (1)$$

  where score_i is 0 if the i-th presentation is classified as genuine, and 1 otherwise. When multiple categories of PAIs exist, we report the overall APCER as the average across all attack categories.

- **BPCER** (*bona-fide* Presentation Classification Error Rate): BPCER measures the proportion of genuine presentations incorrectly classified as attacks. For N_{BF} genuine presentations, BPCER is calculated as:

$$\text{BPCER} = \frac{1}{N_{\text{BF}}} \sum_{i=1}^{N_{\text{BF}}} \text{score}_i. \qquad (2)$$

- **ACER** (Average Classification Error Rate): ACER is the average of APCER and BPCER:

$$\text{ACER} = \frac{\text{APCER} + \text{BPCER}}{2}. \qquad (3)$$

  The Equal Error Rate (EER), used to determine the score threshold on the validation (val) set, approximates the ACER for the validation set, where $\text{APCER}_{\texttt{val}} \approx \text{BPCER}_{\texttt{val}}$.

4.2. Results of PAD Experiments

Results of Model1: In our initial experiment, we trained the entire PAD CNN from scratch using the `train` partition of VRPAD, comprising 588 video samples. With a selection of every 10-th frame, the training dataset encompassed a total of $37k$ frames. Despite unequal volumes of classes (*bona-fide* v/s PA) and various attack instruments, we did not perform any class balancing during data loading or loss calculations. For baseline experiments, we relied on the default setup, although techniques for balancing data distributions could potentially improve PAD performance.

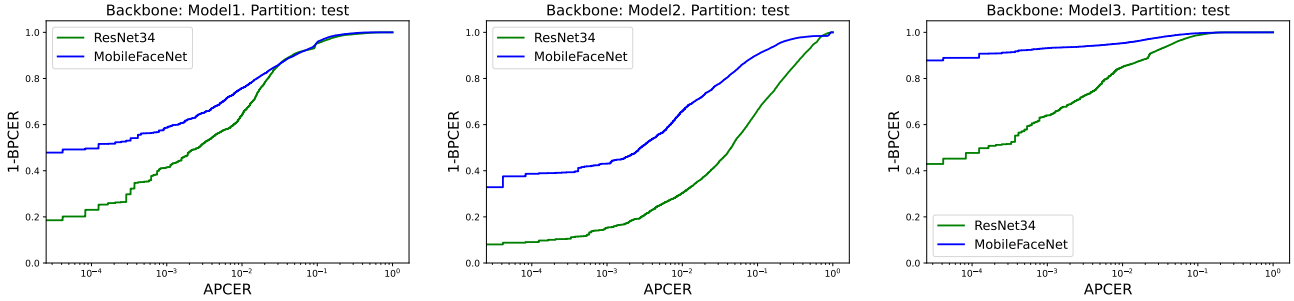
For Model1 with a ResNet-34 backbone, an EER of 10.46% was achieved on the `dev` partition. Applying the same score threshold on the `test` partition an APCER (average) of 7.69% and BPCER of 7.50% were obtained. This indicates overall approximately 6,500 frames out of 43,648 were inaccurately classified by Model1 of the PAD CNN. Similarly, when the PAD CNN with a MobileFaceNet backbone was trained from scratch, a nearly identical EER (10.42%) was observed compared to the ResNet backbone. The APCER and BPCER on the `test` partition were 7.82% and 7.07%, respectively. The results of PAD assessment are presented in Table 2, while Receiver Operating Characteristic (ROC) curves for both backbones are shown in Fig. 5a. Although both backbones exhibited similar performance around the EER operating point (of the validation partition), the ROC analysis reveals relatively superior performance of the MobileFaceNet architecture at lower ranges of APCER. Table 2 also provides a summary of incorrectly classified attack series (such as mannequins or flex masks). It can be observed that MobileFaceNet-based configuration is better at detecting most types of attacks compared to its ResNet-based counterpart.

Results of Model2: In this configuration, the pretrained backbone serves as a fixed feature extractor, and training is restricted to the classifier component only. The ResNet and MobileFaceNet backbones comprise 34,171,713 and 1,233,409 parameters, respectively, while our FF classifier has 32,897 parameters. Thus, in the Model2 setup, only 0.09% (for ResNet) and 2.66% (for MobileFaceNet) of the overall PAD CNN parameters are trainable. For the ResNet backbone, we observed BPCER of 16.44%, along with a notably high APCER of 27.12%, indicating that nearly one in every four attack frames was misclassified as genuine. Using the MobileFaceNet backbone, the error rates were nearly halved compared to the ResNet backbone, although they exhibited similar trends. The subpar performance of the Model2 configuration is further evident from Fig. 5b. This can potentially be attributed to the fact that less than 0.09% of the overall CNN was trainable, which may be insufficient, for the given training setup, to effectively mitigate the threat of presentation attacks on the HMD. Additionally, employing a different backbone, such as one trained on more diverse and generic datasets, may lead to different conclusions.

Results of Model3: This configuration involves overall finetuning of the PAD CNN, where the pretrained backbone is finetuned while the classifier is trained. Thus, it can be anticipated to harness the benefits of pretrained weights (from rich and dynamic data) and adapting the same for the specific task using the appropriate dataset, *i.e.*, the training partition of VRPAD. The PAD assessment provided in Table 2 validates this assumption. In case of the ResNet backbone, the BPCER was significantly reduced to as low

PAD Model	Backbone	EER (dev)	APCER 2	APCER 3	APCER 4	APCER 5	APCER 6	APCER 7	APCER	BPCER	ACER
Train from scratch	ResNet	10.46	2.44	0.42	61.60	13.88	1.50	0.00	7.69	7.50	7.59
	MobileFaceNet	10.42	0.64	0.00	65.00	13.88	2.22	0.00	7.82	7.07	7.45
Fixed backbone	ResNet	32.65	61.03	37.18	56.35	19.30	32.61	7.28	27.12	16.44	21.78
	MobileFaceNet	14.21	43.08	8.08	24.81	20.07	3.16	0.98	12.24	8.14	10.19
Finetune	ResNet	13.65	3.50	0.19	40.51	12.00	19.40	0.83	9.68	1.31	5.50
	MobileFaceNet	8.70	1.28	0.00	18.46	0.95	5.41	0.00	2.57	2.60	2.59

Table 2. Assessment of PAD for VRPAD using three training configurations on both backbones. The EER values are obtained from the `val` partition, while all other metrics are computed on the `test` partition. The interim columns with *APCER x* refer to the APCER obtained for specific attack series (Refer Table 1). All values are indicated as percentages. As each value is an error metric, lower values are desirable.



(a) (b) (c)

Figure 5. ROC (Receiver Operating Characteristics) plots reflecting the performance of PAD using each Model configuration on the `test` partition of the VRPAD.

as 1.31% compared to 7.5% when the entire PAD CNN was trained from scratch. Although there was a marginal increase of nearly 2% each in the APCER (on the test partition) and ACER (on the validation partition) for this backbone, the overall performance on the test partition exhibited improvement. For the MobileFaceNet backbone, Model3 demonstrated the best performance, with error rates on the test partition of VRPAD as low as 2.60%, and both error rates (APCER and BPCER) balanced around the ACER operating point. For each type of attack (denoted by attack series), the MobileFaceNet outperforms the ResNet-based model by a significant margin. The ROC plots in Fig. 5c clearly indicate that Model3, particularly for the MobileFaceNet backbone, offers better capabilities for PA detection.

A simple visual comparison of ROCs from Fig. 5 provides two clear observations: first, Model3 (involving overall finetuning of pretrained weights) outperforms the detection of PAs from VRPAD. Second, across all model configurations, MobileFaceNet is significantly better than ResNet34 for PAD within the given experimental setup.

5. Conclusions

As VR headsets continue to gain traction across diverse applications, it is imperative for research community to anticipate and address the potential risks associated with their use, particularly in use-cases involving user authentication. Despite the awareness of possible risks, the lack of periocular video data captured from VR headsets has hindered the validation and study of these concerns. In this work, we made an important contribution by introducing a new dataset of periocular videos captured from the internal camera of a specific VR device (Meta Quest Pro). This publicly available dataset, VRPAD, comprising 900 *bona-fide* videos from 25 subjects and 996 presentation attacks, enables researchers to comprehensively investigate the risks associated with biometric authentication systems based on VR headsets to various types of PAs. Notably, the *bona-fide* videos are captured under different conditions, including steady gaze, moving gaze, and partially closed eyes, with and without glasses, while the presentation attacks encompass six distinct instruments.

The second contribution of our work involves assessing the performance of two prominent CNN architectures (ResNet34 and MobileFaceNet) using different training configurations on the newly created dataset. Our experiments establish baseline performance metrics and reveal the vulnerability of samples from VR headset to spoofing attacks. Despite achieving acceptable performance in terms of ACER around the specific operating point, we observe a significant drop in performance at lower values of APCER, indicating a significant risk of spoofing attacks. For instance, when the APCER falls below 1%, the BPCER reaches nearly 50% in some experiments– implying that nearly every other *bona-fide* frame is incorrectly classified as an attack.

Our experiments demonstrate that the MobileNet-based architecture consistently outperforms the ResNet-based architecture, indicating the significance of architecture selection in mitigating the PAD threat. Further research is warranted to enhance the accuracy and robustness of PAD methods, particularly in anticipation of potential advancements in spoofing techniques.

As a future work, we aim to explore suitable PAD models and training strategies to detect these attacks effectively. Additionally, developing better methods for detecting or localizing eye or iris regions may lead toward improving PAD performance and adapting existing techniques to VR headset data. Addressing these challenges in a timely manner is crucial to ensuring the security and reliability of VR-based authentication systems in the face of evolving threats.

References

[1] A. Agarwal, A. Noore, M. Vatsa, and R. Singh. Generalized contact lens iris presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):373–385, 2022. 2

[2] A. Agarwal, N. Ratha, A. Noore, R. Singh, and M. Vatsa. Misclassifications of contact lens iris pad algorithms: Is it gender bias or environmental conditions? In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 961–970, 2023. 2

[3] A. Barrett, A. Pack, D. Monteiro, and H.-N. Liang. Exploring the influence of audience familiarity on speaker anxiety and performance in virtual reality and real-life presentation contexts. *Behaviour & Information Technology*, 43(4):787–799, 2024. 1

[4] G. Bente, F. Eschenburg, and N. C. Krämer. Virtual gaze. a pilot study on the effects of computer simulated gaze in avatar-based conversations. In *Virtual Reality: Second International Conference, ICVR 2007*, pages 185–194. Springer, 2007. 1

[5] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer, 2018. 5

[6] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012. 2

[7] P. Dhar, A. Kumar, K. Kaplan, K. Gupta, R. Ranjan, and R. Chellappa. Eyepad++: A distillation-based approach for joint eye authentication and presentation attack detection using periocular images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20218–20227, 2022. 2

[8] I. C. Duta, L. Liu, F. Zhu, and L. Shao. Improved residual networks for image and video recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 9415–9422. IEEE, 2021. 5

[9] M. Fang, F. Boutros, and N. Damer. Intra and cross-spectrum iris presentation attack detection in the nir and visible domains. In *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pages 171–199. Springer, 2023. 2

[10] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019. 2

[11] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019. 2

[12] V. T. C. Koh. Applications of extended reality in ophthalmology: Systematic. *Journal of Medical Internet Research*, 23(8):e24152, 2021. 1

[13] K. Kotwal and S. Marcel. Cnn patch pooling for detecting 3d mask presentation attacks in nir. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1336–1340. IEEE, 2020. 5

[14] Y. Li, Y. Lian, J. Wang, Y. Chen, C. Wang, and S. Pu. Few-shot one-class domain adaptation based on frequency for iris presentation attack detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2480–2484. IEEE, 2022. 2

[15] Y. Li, J. Wang, Y. Chen, D. Xiey, and S. Pu. Single domain dynamic generalization for iris presentation attack detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[16] C. D. McMurrough, V. Metsis, J. Rich, and F. Makedon. An eye tracking dataset for point of gaze detection. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 305–308, 2012. 2

[17] K. Patel, H. Han, and A. K. Jain. Cross-database face antispoofing with robust feature representation. In *Biometric Recognition: 11th Chinese Conference, CCBR 2016, Chengdu, China, October 14-16, 2016, Proceedings 11*, pages 611–619. Springer, 2016. 5

[18] I. Rigas and O. V. Komogortsev. Eye movement-driven defense against iris print-attacks. *Pattern Recognition Letters*, 68:316–326, 2015. 2

[19] R. Sharma and A. Ross. D-netpad: An explainable and interpretable iris presentation attack detector. In *IEEE international joint conference on biometrics (IJCB)*, pages 1–10. IEEE, 2020. 2

[20] V. S. Swarup, D. Sadhya, V. Patel, and K. De. Presentation attack detection in iris recognition through convolution block attention module. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2022. 2

[21] M. Tonsen, X. Zhang, Y. Sugano, and A. Bulling. Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. In *Proceedings of the ninth biennial ACM symposium on eye tracking research & applications*, pages 139–142, 2016. 2

[22] E. Waisberg, J. Ong, M. Masalkhi, N. Zaman, P. Sarker, A. G. Lee, and A. Tavakkoli. The future of ophthalmology

and vision science with the apple vision pro. *Eye*, 38(2):242–243, 2024. 1

[23] S. Yadav and A. Ross. Cit-gan: Cyclic image translation generative adversarial network with application in iris presentation attack detection. In *IEEE/CVF winter conference on applications of computer vision*, pages 2412–2421, 2021. 2

[24] A. W. K. Yeung, A. Tosevska, E. Klager, F. Eibensteiner, D. Laxar, J. Stoyanov, M. Glisic, S. Zeiner, S. T. Kulnik, R. Crutzen, et al. Virtual and augmented reality applications in medicine: analysis of the scientific literature. *Journal of medical internet research*, 23(2):e25499, 2021. 1