Assessing the Performance of Efficient Face Anti-Spoofing Detection Against Physical and Digital Presentation Attacks

Luis S. Luevano, Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Miguel Gonzalez-Mendoza, Davide Frey

Assessing the Performance of Efficient Face Anti-Spoofing Detection Against Physical and Digital Presentation Attacks

Luis S. Luevano[1] Yoanna Martínez-Díaz[2] Heydi Méndez-Vázquez[2]
Miguel González-Mendoza[3] Davide Frey[1]

[1]Univ Rennes, Inria, CNRS, IRISA, France
263 Av. Général Leclerc, 35042, Rennes, France
{luis-santiago.luevano-garcia, davide.frey}@inria.fr
[2]Advanced Technologies Application Center (CENATAV)
7A #21406 Siboney, Playa, P.C.12200, Havana, Cuba
{ymartinez, hmendez}@cenatav.co.cu
[3]School of Engineering and Sciences, Tecnologico de Monterrey
Av. Eugenio Garza Sada 2501 Sur, Monterrey, P.C. 64849, N.L. Mexico
mgonza@tec.mx

Abstract

In this paper, we examine how pre-processing and training methods impact on the performance of Lightweight CNNs through evaluations on MobileNetV3 with a spoofing detection head, dubbed "MobileNetV3-Spoof". Using the UniAttackData dataset from the 5th Face Anti-Spoofing Challenge@CVPR2024, which covers a broad spectrum of spoofing scenarios including deepfake and adversarial attack samples, we assess how well the model performs over different setups, including pre-trained models and models trained from scratch with or without initial face detection and alignment. Our results show that pre-processing steps significantly boost the model's ability to identify spoof samples, especially against complex attacks. Through detailed comparisons, we offer insights that could guide data curation and the creation of more effective and efficient anti-spoofing techniques suitable for real-world use in the era of digital face attacks. We make our code publicly available at: https://github.com/Inria-CENATAV-Tec/Assessing-Efficient-FAS-CVPR2024

1. Introduction

In the rapidly evolving landscape of digital security, facial recognition systems have become a cornerstone technology, employed across a myriad of applications from smartphone unlocking mechanisms to sophisticated border control and surveillance systems. Despite their widespread adoption and the advancements in accuracy and reliability, these systems remain vulnerable to an array of spoofing attacks. These attacks, ranging from simple photograph-based spoofs to more sophisticated modalities, including 3D mask attacks [18], pose a significant threat to the integrity of biometric authentication systems [23]. As such, the development of effective anti-spoofing measures [25] is paramount to safeguarding these technologies against unauthorized access [12].

Among the various strategies employed to enhance the security of facial recognition systems, lightweight Face Anti-Spoofing (FAS) [24, 31] has emerged as a particularly promising avenue of research. These models offer the dual benefits of high efficiency and robust performance, making them ideally suited for real-time processing on devices with limited computational resources. Within this context, MobileNetV3, an effective and effective model for its exceptional balance between accuracy and efficiency, has been previously used for physical FAS [24] with modifications to the classification head. We choose to study this variant, referred as "MobileNetV3-Spoof" in this paper.

We note that the effectiveness of FAS models in real-world scenarios is contingent upon several factors, not least of which is the type of input data. As such previous editions of the of the Face Anti-Spoofing Challenge@CVPR [6, 15, 16, 18, 28] have addressed a variety of scenarios using datasets for large-scale multi-modal FAS [32, 33], cross ethnicity [17], high-fidelity 3D masks [19], and surveillance scenarios [7].

Previous analysis on efficient network architectures [24] consisted in extensively testing different variants of physical attacks [28], while in this work we focus on testing on large-scale physical and digital attacks through the UniAttackData [8] dataset and its three-protocol setup in the context of the 5th Face Anti-Spoofing Challenge@CVPR2024 [27]. We study the role of pre-processing techniques, particularly face detection and alignment, in affecting the model performance of MobileNetV3-Spoof. Facial pre-processing techniques are critical in normalizing input data, ensuring that the model focuses on relevant facial features, and reducing the impact of variations in pose and facial expressions [5]. Moreover, we assess the choice of training data and the approach to model training—ranging from pre-training on generic datasets like ImageNet [4] to more specialized datasets designed for anti-spoofing tasks, which further influence the model's ability to generalize across different types of spoofing attacks.

Through a series of experiments comparing different combinations of pre-training, and fine-tuning with face detection and alignment techniques, we seek to understand how each element contributes to the overall performance of Efficient Face Anti-Spoofing models involving adversarial and deepfake data. This exploration is important for our understanding of Efficient and Lightweight Face Anti-Spoofing technologies and to enhance the security of facial-recognition systems against increasingly sophisticated digital threats.

In this work, we contribute with the following:

- Provide a baseline for physical and digital attack vectors using MobileNetV3-Spoof, a representative backbone for Efficient Face Anti-Spoofing, examining the impact of various pre-processing and training protocols on its effectiveness in detecting spoofing attacks.
- Provide a working experimentation methodology for testing and achieving results in the UniAttackDataset, representing a wide range of both physical and digital spoofing attacks.[1]
- Provide specific insights for improving performance on Efficient CNN deployments for Face Anti-Spoofing in practice.

This paper is organized as follows: Section 2 explains the related work relevant for our approach, Section 3 explains our experimentation methodology, Section 4 shows our achieved results in Phase 1 and Phase 2 of the 5th Face Anti-Spoofing Challenge@CVPR2024 competition [27], Section 5 presents our insights regarding the results obtained in the competition, and Section 6 gives our closing remarks for this study.

[1]Our code is available at https://github.com/Inria-CENATAV-Tec/Assessing-Efficient-FAS-CVPR2024

2. Related work

In this section, we present recent work tackling to the Face Anti-Spoofing (FAS) problem. We cover recent methods based on Vision Transformers (ViTs) and Efficient Convolutional Neural Networks (CNNs).

2.1. Vision Transformers (ViTs) for FAS

ViTs, known for their remarkable performance in extracting global representations, have also been a popular choice for FAS approaches. We explore popular methods in this section.

In Class Free Promt Learning (CFPL) [21] the authors used separate encoders for the image and the text labels using two transformer architectures, and learn two separate semantic representations. They optimize cross-dataset generalization using text supervision, diversification of style prompt, and prompt modulation.

The Multi-domain Incremental Learning approach [29] utilizes three blocks: the Active Domain Experts (ADE) Block, the Instance-Wise Router (IwR), and the Asymmetric Classifier. The ADE blocks are ViT-inspired modules separating parameters domain-invariant and domain-specific sets of weights and chooses either at inference time by using the IwR to map a domain center a training time and measuring the sample's cosine similarity to it afterwards. The Asymmetric Classifier compensates for the different domains from the diverse spoof sample class centers.

The Flexible Modal ViT (FM-ViT) [20] uses a framework comprised of a Multi-Modal Tokenization Module with Cross-Modal Transformer Blocks for image patch encoding and two separate attention mechanisms. This dual branch approach is designed to enhance recognition from a single modal data using multi-modal data.

In Modality-Agnostic ViT (MA-ViT) [14], the authors proposed to eliminate modality-related information to achieve better generalization. Their approach consists on using a novel Modality-Agnostic Transformer Block (MATB) with two attention module designs to promote the discrimination of modality-irrelated patch tokens from inner-modal and inter-modal information.

2.2. Efficient CNNs for FAS

In the evolving landscape of efficient Face Anti-Spoofing technologies [31], the development of neural network architectures that balance computational efficiency with high performance has been pivotal for deployment in real-world applications. Studies involving traditional Lighweight CNN backbones have appeared [24], discussing the performance of efficient backbones such as ShuffleNetv2 [22], MicroNet [13], and MobileNetV3 [11] for the large-scale FAS task. The authors highlighted the effectiveness of the MobileNetV3 backbone, modified to address the unique challenges of Face Anti-Spoofing (called MobileNetV3-Spoof

in this paper). This MobileNetV3-Spoof variant adapts this architecture specifically for the task of distinguishing genuine facial presentations from fake ones by using a spoof-detection head.

2.3. Self-Supervision and Face Detection

In parallel to these developments, the exploration of Self-Supervised methods [3, 9] has introduced novel approaches to learning robust feature representations without extensive labeled datasets. The work encapsulated in Simple Siamese Self-Supervision [3] exemplifies this trend, proposing a methodology that leverages the inherent structure of unlabeled data to improve model performance. While this approach has shown promise across several domains, its application within the context of Efficient Face Anti-Spoofing represents an intriguing frontier for research.

Complementing the advancements in model architecture and self-supervision, significant strides have been made in the realm of face detection methods, such as the RetinaFace detector [5], which provides a competent backbone for the initial stages of anti-spoofing pipelines, ensuring that subsequent analyses are based on well-aligned and correctly identified facial regions.

3. Methodology

Our methodology is designed to systematically evaluate the impact of pre-processing and fine-tuning strategies on the performance of MobileNetV3-Spoof. By focusing on these aspects, we aim to assess the model's effectiveness when testing it under the diverse and challenging conditions presented by the UniAttackData dataset.

3.1. MobileNetV3-Spoof Backbone

This approach, explored in [24], extends the MobileNetV3 backbone for the Face Anti-Spoofing scenario. MobileNetV3 uses a platform-aware Network Architecture Search (NAS) approach for finding optimal global network structures along with the NetAdapt algorithm to search for the number of filters per layer. Additionally, it uses Squeeze-And-Excitation blocks on residual layers and includes a more efficient embedding-extraction structure compared to MobileNetV2 [26]. The spoofing classification head is comprised of Dropout, BatchNorm, Swish, and Linear layers with a Softmax layer when producing the final predictions.

3.2. Face Anti-Spoofing Datasets

In this section, we describe the datasets included in our scenarios. These datasets are used in different training scenarios with testing on the dev and test sets of UniAttackData.

FAS Challenge@CVPR2023-Wild Track dataset The dataset used for the 4th Face Anti-Spoofing Challenge@CVPR2023-Wild Track [28], is a large-scale in-the-wild collection tailored for physical Face Anti-Spoofing. It features 529,571 authentic images representing 148,169 distinct identities, with 853,729 fake face images across 300K identities. The dataset categorizes spoof attempts into three primary types: 2D Print, 2D Display, and 3D PAS, further divided into 17 subcategories, offering a broad spectrum of scenarios for robust anti-spoofing research.

UniAttackData We employ the UniAttackData [8] from the Face Anti-Spoofing Challenge CVPR2024 [27] for testing. This dataset includes three protocols with their own training, development, and test sets. Protocol 1 (p1) is comprised of physical and digital images in training, development, and testing sets. Protocol 2.1 (p2.1) includes training and development data with live probes and digital attacks, excluding physical attacks, while the test set excludes digital attacks and includes physical attacks. Protocol 2.2 (p2.2) has live probes and physical attacks with no digital attacks in training and development sets but makes digital attacks present in its test set without physical attacks. Its attack distribution ranges from 6 types of Adversarial attacks, 6 types of DeepFake attacks, and physical attacks from the CASIA-SURF CeFa [17]. The total number of images in the dataset is 2.5Million with 1800 identities.

Figure 1. Examples of live images and spoof attacks from the UniAttackData database. Spoof attack vectors include physical and digital samples.

3.3. Pre-processing for UniAttackData

The pre-processing pipeline is a critical component of our approach, aimed at standardizing input data to optimize the performance of the Face Anti-Spoofing model. This process involves two main steps: face detection and face alignment, which are executed sequentially on each image before it is fed into the MobileNetV3-Spoof model.

Face Detection: We employ the ResNet50-RetinaFace detector, a state-of-the-art face-detection model known for

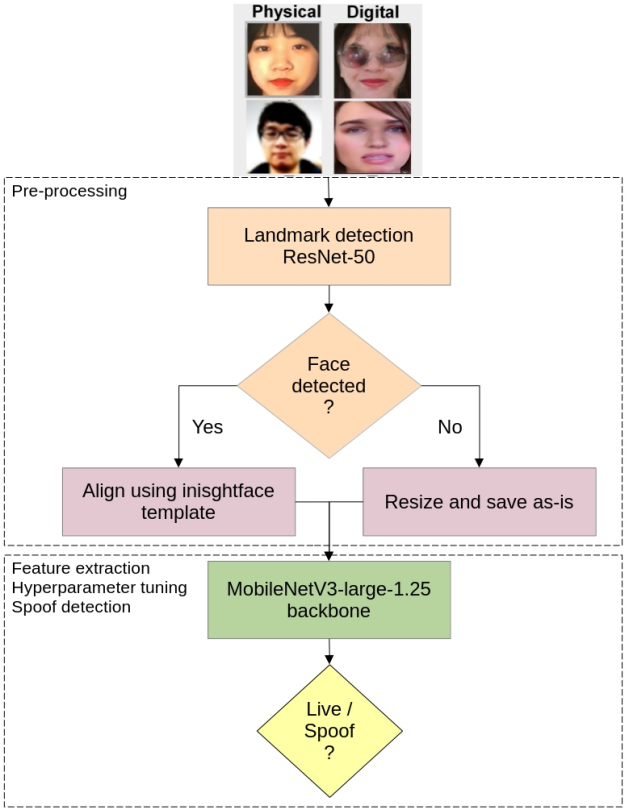

Figure 2. Pre-processing approach for the UniAttackData dataset.

its high accuracy and efficiency in detecting faces across a wide range of poses and lighting conditions. This model leverages the power of ResNet50's deep residual learning framework, combined with RetinaFace's focus on capturing facial landmarks. This model detects 5 landmarks that can be used for posterior processing.

Face Alignment: Once a face is detected, we use the 5 landmarks from the RetinaFace detector to align the image to the Insightface frontal template [10]. This alignment process involves adjusting the detected faces to a standard pose and scale. This step is crucial for reducing variability in the input data due to pose and expression differences, thereby enabling the MobileNetV3-Spoof model to focus more on the distinguishing features relevant to spoof detection. We warp the face to the 128×128 template size.

If no face is detected in an image, the image is resized as-is, without any alignment. This ensures that our model remains robust to scenarios where face detection may fail, by allowing the network to still attempt a classification based on the available visual information. This workflow is illustrated in Figure 2

Augmentation Firstly, we up-scale the image to a 224×224 resolution using bicubic interpolation. We apply random augmentations such as horizontal flip, ISO Noise simulation, random brightness and contrast, and motion blur simulation using the Albumentations library [2]. Finally, we normalize the samples with the training data's mean and standard deviation.

3.4. Training strategies

To test the MobileNetV3-Spoof model within different circumstances during the Face Anti-Spoofing Challenge at CVPR2024 [27] using the UniAttackData dataset, we employed fine-tuning and training from scratch processes involving several distinct training configurations. These configurations were designed to explore the model's adaptability and performance under different initial conditions and training settings.

3.4.1 Phase 1

In this section, we explore our baselines for Phase 1 of the Challenge. The evaluation of this phase includes only the development set of the UniAttackData dataset.

Pre-training with ImageNet and Fine-tuning Initially, the model is pre-trained on the ImageNet dataset, a diverse collection of images spanning a wide range of categories. This pre-training phase provides a solid foundation of general visual features. Subsequently, the model undergoes fine-tuning on the UniAttackData development set, allowing it to adjust its weights to better recognize the specific characteristics of spoofing attacks.

Pre-training with Wild Face Anti-Spoofing and No Fine-tuning: In this configuration, the model is pre-trained on the Wild Face Anti-Spoofing dataset from the CVPR2023 Face Anti-Spoofing challenge. This dataset is specifically designed for anti-spoofing but without its digital counterparts. Unlike the previous approach, this model is not fine-tuned further, testing its ability to generalize from the training alone on extensive spoofing data.

3.4.2 Phase 2

In this section, we explore our final approaches for Phase 2 of the competition, which uses the development and test sets of the UniAttackData dataset.

Pre-training with Simple Siamese Self-Supervision and Fine-tuning: Leveraging the principles of self-supervised learning, this approach involves pre-training the model using a Simple Siamese network architecture designed for self-supervision without labels. This pre-training setting is

aimed at enabling the model to learn robust feature representations without reliance on labeled data. Following this, the model is fine-tuned on the UniAttackData development set.

Training from scratch: We also explore the performance of MobileNetV3-Spoof models trained from scratch on the UniAttackData development set, both with and without the Face Detection and Alignment pre-processing steps described in subsection 3.3. This comparison serves to underscore the impact of pre-processing and pre-training on the model's spoof detection capabilities.

3.5. Experimental environment

Our backbone is the MobileNetV3 of 1.25 width and a Live/Spoof detection head, as employed in [24]. Our hyperparameters for training from scratch include a multi-step learning rate starting at 0.1 multiplied by 0.1 at steps 200 and 400 using an SGD optimizer, "relu" activation function, a 1280 feature vector, and up to 500 epochs. For fine-tuning, we employ a learning rate starting at 0.001 with the same reduction steps. [2]

For the Simple Siamese with Self-Supervision [3], we used the implementation present in [30].

3.6. Metrics

In this section, we describe the metrics for evaluating the performance of our approaches. The Attack Presentation Classification Error Rate (APCER) in Eq. 1 measures the rate at which spoofing attacks are incorrectly classified as genuine access attempts, providing insight into the model's vulnerability to false negatives.

$$APCER = \frac{FP}{TN + FP} \tag{1}$$

Conversely, the Bona Fide Presentation Classification Error Rate (BPCER) in Eq. 2 assesses the frequency with which genuine presentations are misidentified as spoofs, highlighting potential issues with false positives.

$$BPCER = \frac{FN}{FN + TP} \tag{2}$$

The Average Classification Error Rate (ACER) (Eq. 3 serves as a balanced metric, averaging the APCER and BPCER to offer a holistic view of overall performance.

$$ACER = \frac{APCER + BPCER}{2} \tag{3}$$

Additionally, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve further encapsulates the model's discriminative power, reflecting its ability to distinguish between genuine and spoofed presentations across varying thresholds.

[2]Our code is available at https://github.com/Inria-CENATAV-Tec/Assessing-Efficient-FAS-CVPR2024

4. Results

In this section, we report the results on the dev set of UniAttackData for Phase 1 of the challenge, and for its test set on Phase 2.

4.1. Phase 1

For our training methodologies with pre-training using Imagenet with fine-tuning (Section 3.4.1) and FAS@CVPR2023-Wild without fine-tuning (Section 3.4.1), we present Table 1 with our results. We note specially note the superior performance of fine-tuning over the UniAttackDataset. Even when the FAS@CVPR2023-Wild dataset contains a large amount of spoof images from physical sources, it is not sufficient for achieving the best performance over physical and digital data.

Pre-training dataset	Fine-tuning	APCER (%)	BPCER (%)	ACER (%)	AUC (%)
FAS@CVPR2023	No	37.00	37.00	37.00	66.33
ImageNet	Yes	**6.39**	**6.38**	**6.39**	**96.61**

Table 1. Results on the dev set of the UniAttackData for Phase 1 of the competition with the MobileNetV3-Spoof backbone.

4.2. Phase 2

This phase of the challenge included dev and test data from the UniAttackData dataset. We present the achieved results over the naïve Simple Siamese (SimSiam) Self-Supervised setting with fine-tuning (Section 3.4.2) and training from scratch (no pre-training) with and without our face alignment approach (Section 3.4.2). Our results are compiled in Table 2.

Training method	APCER (%)	BPCER (%)	ACER (%)	AUC (%)
SimSiam-aligned-FT	33.82	57.21	45.52	54.57
Scratch-no-align	**4.37**	66.42	35.39	77.18
Scratch-aligned	25.14	**0.91**	**13.02**	**99.49**

Table 2. Results on the test set of the UniAttackData for Phase 2 of the competition with the MobileNetV3-Spoof backbone. FT denotes fine-funing on the Simple Siamese pre-trained approach.

In these settings, the Simple Siamese naïve strategy did not perform as expected for the UniAttackData test set, specially noting the BPCER for the live labels, even when fine-tuning. On the other hand, when training from scratch, the

results are more favorable. Particularly, with our face detection and alignment pre-processing pipeline, we note a specially lower ACER in this scenario. Breaking down the ACER metric, we observe that our approach favors the detection of live subjects. In comparison, the Attack Presentation Error is higher in this case. We note that the opposite is true when we do not align the images, where the spoof attack labels are more easily detected by the algorithm.

5. Discussion

In this section, we will cover our insights for the performance of our approach and point out the limitations of our study with potential avenues applicable to efficient approaches in testing the performance on physical and digital attack scenarios.

5.1. Performance

We break down our performance discussion with the resulting accuracy of our approach for lightweight CNNs on the UniAttackData dataset and comment on the efficiency vs accuracy trade-off for this approach.

On Accuracy performance for UniAttackData With our results from Phase 1, we still notice a clear gap for dataset generalization using lightweight approaches and the regular pipeline of face detection, face normalization, and Face Anti-Spoofing detection. Without fine-tuning for the specific data distribution, it becomes increasingly challenging for only a lightweight algorithm to generalize for unseen attacks. This is also extended for the digital scenario, where we would need additional synthetic data or adversarial training methodologies to improve cross-dataset performance.

For our test-set scenario, we note three phenomenons affecting performance on this dataset: an increased number of samples from the spoof label in protocols p1 and p2.1, the effect of alignment when training from scratch, and the capability of the network to adjust to live labels. Self-supervision appears to be more sensitive to the dataset imbalance, with a better performance on detecting physical and digital attacks. At the same time, we notice a major difference in the APCER and BPCER when removing the alignment, with the highest error inverting on each scenario. In our experiments, we noticed more images from the spoof labels with heavy distortion artifacts, skipping the alignment face. In addition, we noticed that the original UniAttackData samples from the "live" class label are not cropped focusing on the face image only. This means that they contain mid or full-body shots and objects from the background scenery, which proved to be detrimental for model performance when no additional pre-processing is done. This is why our face detection and alignment pre-processing significantly helped to improve model performance for this dataset and class label. This is particularly demonstrated by the loss of BPCER performance with the non-aligned version of the dataset.

Efficiency vs Accuracy trade-off on MobileNetV3-Spoof and UniAttackData The complexity of the MobileNetV3-Spoof backbone is 0.38 GFLOPs with 4.75 Params, being an extremely lightweight proposal for deployment on real-time scenarios on constrained hardware. The result of the scratch training with aligned samples is the most effective by the BPCER, AUC, and ACER metrics, where the ACER metric under 0.4GFLOPs makes it specially effective. Bridging the accuracy gap with the APCER performance of the non-aligned version, and the BPCER of the aligned version, would make this implementation one of the most compelling use cases for lightweight and highly accurate physical and digital presentation attack detection.

5.2. Limitations and future work

In this section, we explore the limitations of our approach, which open several avenues for future research and development. We share our insights that could assist in improving the performance of efficient anti-spoofing solutions given the physical and digital attack vectors present in this study.

One of the primary limitations observed is the challenge in generalizing the performance of lightweight models across unseen data distributions without specific fine-tuning. This limitation is particularly pronounced in scenarios involving digital attacks, underscoring the need for more sophisticated training methodologies. Our findings indicate a sensitivity to dataset imbalance, especially in the context of self-supervision. The variation in performance metrics such as APCER and BPCER between aligned and non-aligned training scenarios highlights the model's dependency on the quality of input data pre-processing. The presence of distortion artifacts in digital spoof images, obscured subjects on physical attack data, and the inclusion of additional objects in live class images point to the necessity of robust face detection and alignment processes. The standardization provided by these pre-processing steps is crucial for the success of lightweight algorithms, as evidenced by the observed performance gaps.

To address the issue of dataset generalization and improve the model's robustness against digital spoofing attacks, future work could explore the incorporation of adversarial training techniques and the utilization of synthetic data. These methodologies could improve the model's ability to learn more generalized features that are effective across diverse spoofing scenarios. Given the sensitivity to dataset imbalance, there is potential to refine self-supervision techniques to better accommodate the characteristics of Face Anti-Spoofing datasets. This could involve developing more balanced self-supervised learning

approaches that are less susceptible to the distribution of labels within the training data coupled with the regular Softmax loss function. Using simultaneously the Softmax loss aided by metric learning losses with different views, as in Self-Supervised approaches, could further improve the robustness of lightweight methods. Building on the efficiency vs accuracy trade-off highlighted by our study, future efforts could focus on optimizing lightweight models to further improve the gap between these two critical aspects. In face recognition, for example, approaches like GhostFaceNet [1] have achieved a remarkable balance between accuracy and efficiency with face data.

6. Conclusion

In Our exploration into the efficacy of MobileNetV3-Spoof with pre-processing and fine-tuning strategies has unveiled critical insights into developing lightweight Face Anti-Spoofing solutions for physical and digital face spoofing scenarios. We highlight the importance of pre-processing, specifically face detection and alignment, in influencing the model's ability to distinguish between genuine and spoofed facial presentations. We provide strategies for assessing the detection of physical and digital spoofing attacks and point out potential strategies for optimizing the training of efficient models to achieve significant advancements in detection accuracy, even with limited computational resources. Our findings also underscore the complex interplay between dataset sensitivity, the necessity for pre-processing techniques, and the strategic selection of training methodologies to overcome the inherent challenges in dataset generalization and the efficient detection of sophisticated spoofing attacks. The efficiency vs accuracy trade-off presented by MobileNetV3-Spoof emphasizes the critical need for ongoing research to refine efficient models for real-world applications considering a diverse array of spoofing tactics. We hope the insights from this study serve as a baseline for future efficient models exploring physical and digital attack vectors.

Acknowledgements

This work was partially funded by the SOTERIA H2020 project. SOTERIA received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No101018342. This content reflects only the author's view. The European Agency is not responsible for any use that may be made of the information it contains.

The authors would like to thank the financial support from Tecnologico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID # E120 - EIC-GI06 - B-T3 - D

References

[1] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446, 2023. 7

[2] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 4

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021. 3, 5

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020. 2, 3

[6] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, and Zhen Lei. Surveillance face presentation attack detection challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6360–6370, 2023. 1

[7] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z Li, and Zhen Lei. Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 1

[8] Hao Fang, Ajian Liu, Haocheng Yuan, Junze Zheng, Dingheng Zeng, Yanhong Liu, Jiankang Deng, Sergio Escalera, Xiaoming Liu, Jun Wan, and Zhen Lei. Unified physical-digital face attack detection, 2024. 2, 3

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 3

[10] Jia Guo, Jiankang Deng, Xiang An, Jack Yu, and Baris Gecer. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, 2024. 4

[11] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. 2

[12] Alain Komaty, Vedrana Krivokuća Hahn, Christophe Ecabert, and Sébastien Marcel. Can personalised hygienic masks be used to attack face recognition systems? In *2023*

IEEE International Joint Conference on Biometrics (IJCB), pages 1–10, 2023. 1

[13] Yunsheng Li, Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Lu Yuan, Zicheng Liu, Lei Zhang, and Nuno Vasconcelos. Micronet: Improving image recognition with extremely low flops. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 468–477, 2021. 2

[14] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1180–1186, 2022. 2

[15] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–10, 2019. 1

[16] Ajian Liu, Xuan Li, Jun Wan, Yanyan Liang, Sergio Escalera, Hugo Jair Escalante, Meysam Madadi, Yi Jin, Zhuoyuan Wu, Xiaogang Yu, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biometrics*, 10(1):24–43, 2021. 1

[17] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 1, 3

[18] Ajian Liu, Chenxu Zhao, Zitong Yu, Anyang Su, Xing Liu, Zijian Kong, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. 3d high-fidelity mask face presentation attack detection challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 814–823, 2021. 1

[19] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022. 1

[20] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang Zhen Lei, Du Zhang, Stan Z Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 2

[21] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[22] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Computer Vision – ECCV 2018*, pages 122–138, Cham, 2018. Springer International Publishing. 2

[23] Yoanna Martinez-Diaz, Miguel Nicolas-Diaz, Heydi Mendez-Vazquez, Luis S Luevano, Leonardo Chang, Miguel Gonzalez-Mendoza, and Luis Enrique Sucar. Benchmarking lightweight face architectures on specific face recognition scenarios. *Artificial Intelligence Review*, pages 1–44, 2021. 1

[24] Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Luis S. Luevano, and Miguel Gonzalez-Mendoza. Exploring the effectiveness of lightweight architectures for face anti-spoofing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6392–6402, 2023. 1, 2, 3, 5

[25] Zuheng Ming, Muriel Visani, Muhammad Muzzamil Luqman, and Jean-Christophe Burie. A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices. *Journal of Imaging*, 6(12), 2020. 1

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 3

[27] Jun Wan, Ajian Liu, Jiankang Deng, Shengjin Wang, Ya-Li Li, Sergio Escalera, Hugo J. Escalante, Isabelle Guyon, and Zhen Lei. 5th chalearn face anti-spoofing workshop and challenge@cvpr2024. https://codalab.lisn.upsaclay.fr/competitions/17490, 2024. 2, 3, 4

[28] Dong Wang, Jia Guo, Qiqi Shao, Haochi He, Zhian Chen, Chuanbao Xiao, Ajian Liu, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, Jun Wan, and Jiankang Deng. Wild face anti-spoofing challenge 2023: Benchmark and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 6380–6391, 2023. 1, 2, 3

[29] Keyao Wang, Guosheng Zhang, Haixiao Yue, Ajian Liu, Gang Zhang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Multi-domain incremental learning for face presentation attack detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5499–5507, 2024. 2

[30] Phil Wang. Bootstrap your own latent (byol), in pytorch. https://github.com/lucidrains/byol-pytorch, 2024. 5

[31] Zitong Yu, Yunxiao Qin, Xiaqing Xu, Chenxu Zhao, Zezheng Wang, Zhen Lei, and Guoying Zhao. Auto-fas: Searching lightweight networks for face anti-spoofing. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 996–1000, 2020. 1, 2

[32] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–928, 2019. 1

[33] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 1