IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 February 2024, accepted 28 March 2024, date of publication 4 April 2024, date of current version 17 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3385373

RESEARCH ARTICLE

Securing Phygital Gameplay: Strategies for Video-Replay Spoofing Detection

VIKTOR DÉNES HUSZÁR[1,2], (Member, IEEE), AND VAMSI KIRAN ADHIKARLA[1]
[1]Computer Vision and Artificial Intelligence Team, International Federation of Teqball, 1101 Budapest, Hungary
[2]Faculty of Military Science and Officer Training, National University of Public Service, 1083 Budapest, Hungary

Corresponding author: Vamsi Kiran Adhikarla (vamsi.kiran@fiteq.org)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Physical Virtual Sports (PVS) utilize digital technologies for the analysis and evaluation of sports performances. This research article addresses the challenge of detecting video-replay spoofing in PVS, with a specific focus on a digital football sport aimed at assessing and improving a player's football juggling skills. In the context of the growing presence of digital coaches as well as PVS, accurate assessment of player performance and identification of deceptive practices in these applications are paramount. The proliferation of sophisticated technologies, such as deepfake algorithms and computer vision techniques, has facilitated the manipulation of video replays, deceiving both viewers and officials. To tackle the challenges associated with video-replay spoofing, this article introduces a meticulously curated dataset comprising 600 players engaged in the digital football sport. Additionally, the dataset includes video-replay spoofing videos captured on a wide range of display devices. A deep learning-based model is developed and trained on this dataset, achieving an accuracy rate of approximately 95%. Generalization studies were also conducted to assess the model's ability to generalize to unseen scenarios and datasets. The ROC-AUC score highlighted the model's discriminative power across different threshold values, validating its effectiveness in distinguishing between genuine and spoofed video replays. The results demonstrate that our trained model exhibited consistent performance across multiple public face biometric spoofing datasets, underscoring its robustness against sophisticated video-replay attacks in various domains. Additionally, ablation studies were carried out by systematically removing or modifying the model's backbone architectures to analyze their effects on detection accuracy and reliability. Furthermore, computational complexity analysis was presented to evaluate the model's efficiency in terms of time and space requirements. The findings underscore the scientific significance and relevance of video replay spoof detection in PVS. By presenting a novel dataset (https://www.fiteq.org/research) and employing an advanced deep learning approach, this article contributes to the scientific community's understanding and progress in combating fraudulent practices, ultimately preserving the integrity and fairness of digital sports applications.

INDEX TERMS Active virtual sports, computer vision, dataset, deepfake detection, deep learning, deceptive practices, digital sports applications, fraudulent practices, integrity, video-replay spoofing.

I. INTRODUCTION

In today's diverse world of games, the lines between physical and digital games are getting fuzzy. The Olympic Virtual Series (OVS) signifies this trend, with the International Olympic Committee (IOC) acknowledging "virtual games" at the 9th Olympic Summit [1]. These games exist in two forms: physical and non-physical. Physical virtual games involve players performing physical activities while their data is digitally encoded, such as in Just Dance and indoor cycling. Conversely, non-physical virtual games resemble traditional computer games.

The associate editor coordinating the review of this manuscript and approving it for publication was Xinfeng Zhang.

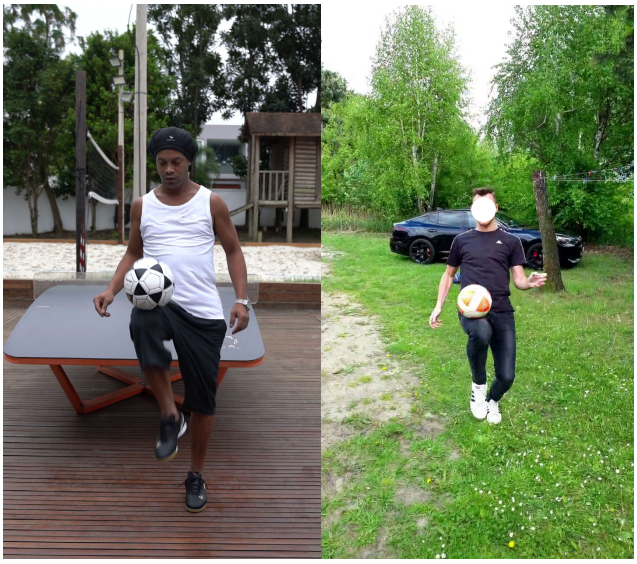

FIGURE 1. Illustration of a phyigital football game scenario in the Sqiller smartphone application: On the left, an expert player like Ronaldinho demonstrates the soccer trick, while on the right, a player is attempting to perform the shown trick. The performance evaluation uses advanced computer vision and deep learning-based techniques.

Currently, physical virtual games in OVS necessitate players' physical presence, posing logistical challenges for global participation. However, advancements in technology make the global virtualization of gaming experiences increasingly feasible. We term these games "phygital games," where remote player participation is facilitated. Phygital games offer more than convenience; they serve as a solution during challenging times like the COVID-19 pandemic, enabling sportsmanship despite physical barriers. By leveraging motion tracking and other technologies, phygital games provide detailed insights into athletes' movements, techniques, and performance, fostering deeper understanding and objective evaluation.

One example of phygital gaming is Sqiller, a smartphone application for football enthusiasts [2]. Users receive instructions for executing soccer tricks through videos featuring expert players like Ronaldinho. The app uses the smartphone's camera, advanced computer vision, and deep learning techniques to monitor and evaluate users' performances compared to the experts'.

The significance of phygital games extends beyond beyond leisure, permeating competitive arenas where global athletes engage in virtual events, enabled by digital technologies for remote competition. Ensuring the integrity of virtual competitions is paramount as phygital games gain traction. Imagine a player's remarkable performance, not from dedication, but shrewd video editing. This digital deception parallels concerns with performance-enhancing drugs, prompting contemplation of a digital equivalent of the World Anti-Doping Agency (WADA) to safeguard fairness. Video-replay spoofing, manipulating video data to deceive viewers, officials, or systems during digital analysis, poses a significant challenge. Advanced digital tools make creating convincing manipulated videos effortless, raising concerns about fairness in phygital games (see Figure 2).

Traditional and deep-learning based video analysis methods also face limitations in phygital gaming scenarios. Techniques like frame-by-frame analysis struggle with subtle alterations or irregularities introduced in video replays, compounded by the dynamic nature of phygital games with evolving environments and player movements.

The specific contributions from the current work to spoof detection for phygital games are:

- A lightweight, deep learning architecture tailored for phygital sports, achieving about 95% accuracy rate on our dataset. Unlike conventional methods for video-replay spoof detection in Human Activity Recognition (HAR) applications, our method adopts a one-step approach for spoof detection, eliminating the need to extract specific user-related facial information as an initial step for classification.
- An efficient strategy to systematically integrate the proposed deep learning network into video capture and live video streams, enabling detection of video replay spoof attacks within phygital games. This novel strategy is designed to seamlessly integrate with the gaming logic responsible for evaluating game performance, enabling real-time execution and eliminating latency concerns, offering a practical and immediate solution for detecting video replay spoof attacks in real-time within phygital games
- A diverse and comprehensive database, comprising real and spoof videos captured from 600 different athletes across various geographical locations, under different lighting conditions, and in varied environments. Our dataset is uniquely diverse, incorporating various capturing and replaying devices, and notably, it explicitly includes videos featuring display bezels. This inclusion addresses a gap in existing video-replay datasets, enhancing the comprehensiveness and applicability of our model in phygital sports scenarios.
- Additional evaluation of the performance of the proposed approach within face biometric recognition applications, assessing its effectiveness beyond the context of phygital games for broader applicability.

The remaining sections of the paper are structured as follows: Section II discusses related work, providing an overview of existing research in the field. Section III presents detailed information on our dataset, along with comparisons to other datasets relevant to video-replay spoof detection. In Section IV, we elaborate on the proposed approaches, including the training methodology employed. Section V presents the findings of our experiments with various deep learning architectures. Within this section, we also examine the generalization potential of our approach and conduct computational complexity analysis. Furthermore, we outline the limitations of our approach. Finally, Section VI

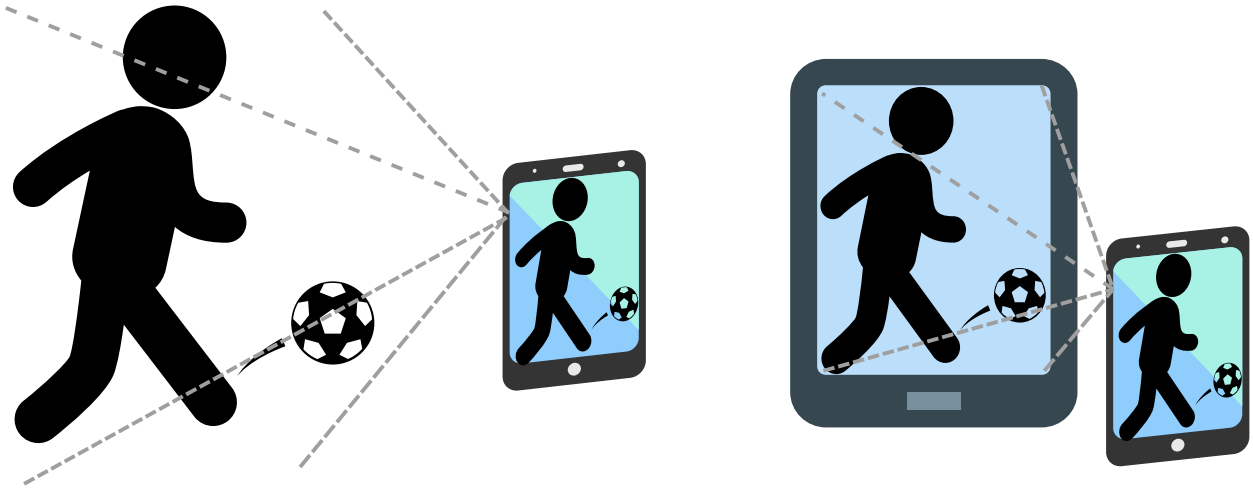

FIGURE 2. Detection of ball juggling events in a hybrid physical-digital (phygital) football game environment–On the left, a legitimate user actively engages in the phygital football game, subject to analysis by a smartphone application equipped with an integrated digital camera. This application utilizes AI-based motion analysis to assess performance. On the right, a potential malicious user attempts to mimic the phygital gaming experience by presenting a prerecorded video replay on a digital screen, as opposed to actively participating in the physical drill'.

summarizes the key findings of the study and outlines potential avenues for future research.

II. RELATED WORK

In the domain of detecting replay attacks from visual data, a wide array of techniques has been developed and explored. However, a predominant focus lies in spoofing detection within biometric recognition applications, particularly involving facial image analysis. These approaches are typically tailored for scenarios where users are positioned close to the camera and are directly facing it.

However, when it comes to the context of phygital games, the dynamics of spoof detection change significantly. In phygital games, athletes are often located at a distance from the camera, and their gaze may not always be directed towards it. As a result, the conventional spoof detection methods that rely on facial image analysis may not be directly applicable to these settings.

Despite the growing interest in phygital games and the increasing popularity of phygital games, there remains a gap in the research concerning spoof detection from videos within this specific context. The lack of extensive work exploring the challenges posed by phygital gaming settings calls for a dedicated investigation to address the unique aspects of video replay spoof detection in phygital games.

In the following section, we provide an overview of related works that have been developed to tackle video replay spoof detection, considering both the existing techniques in biometric recognition and other relevant studies. The examination of these works aims to shed light on the advancements made in this domain and identify potential strategies that can be adapted to suit the specific challenges posed by phygital games.

A. TEXTURE CUE-BASED METHODS

Throughout the literature, texture cue-based methods are extensively utilized in biometric spoof detection applications for detecting face presentation attacks [3]. These methods leverage texture properties to discern between genuine faces and manipulated data, effectively detecting various known attacks, including photo-based, video replay, and 3D mask attacks [3], [4]. In 2010, Tan et al. proposed an early static texture-based method [4] that employed Lambertian models to represent the reflectivities of genuine (alive) faces and face-printed photos. They used the Difference of Gaussian (DoG) filtering technique to derive latent samples [5]. The underlying concept was that a face-printed photo tends to exhibit more distortion than an image of a genuine face, as it undergoes two captures and one print, while genuine faces are only captured once by the biometric system [4]. The method showed promising results with classifiers like Sparse Nonlinear Logistic Regression (SNLR) and Support Vector Machines (SVMs).

In their study, Smith et al. [6] introduced an innovative approach to counterattacks on face recognition systems. Their method involved analyzing the colour reflected from the user's face as displayed on mobile devices. By examining the presence or absence of these reflections, the algorithm can determine whether the images were captured in real time. Notably, the detection of presentation attacks is achieved using straightforward RGB images. To address sensitivity to illumination variations and partial occlusion, Zuiderveld proposed Contrast-Limited Adaptive Histogram Equalization (CLAHE) [7] for image pre-processing [8]. They demonstrated that CLAHE outperformed simple histogram equalization, effectively enhancing the method's performance. Following a similar approach, Bai et al. analyzed micro-textures using Bidirectional Reflectance

Distribution Functions (BRDF) [9]. They extracted the normalized specular component, referred to as the specular ratio image, and computed its gradient histogram, known as the specular gradient histogram. These histograms exhibited distinct shapes for genuine faces and printed photos, enabling the training of a Support Vector Machine (SVM) for face presentation attack detection [9].

Local Binary Pattern (LBP) has also been widely employed as a hand-crafted texture feature in various face analysis-related problems [10]. In 2011, Määttä et al. proposed applying multi-scale LBP to Face Presentation Attack Detection (PAD) [10]. Unlike previous static texture-based approaches, LBP-based methods do not rely on any physical model; instead, they assume that the differences in surface properties and light reflection between genuine faces and planar attacks can be captured by the LBP features. Määttä et al. utilized different LBP configurations, obtaining histograms that were later concatenated to form a global micro-texture feature. This feature was then fed to a non-linear (RBF) SVM classifier for face presentation attack detection [10].

In 2012, Määttä et al. extended their work by incorporating Gabor wavelets and Histogram of Oriented Gradients (HOG) into their framework [11]. These features aimed to capture both facial macroscopic information and facial edges or gradient structures, respectively. The authors used a fast linear SVM and employed late fusion between the outputs of the three SVMs to generate the final decision [11]. Inspired by the context surrounding the face, Yang et al. and Bai et al. proposed using the upper-body region to detect spoofing attacks [9], [12]. Yang et al. employed a 1.6× enlarged face region (H-Face) and segmented canonical facial regions to extract texture features from different components of the face. This approach utilized Local Binary Pattern (LBP), Histogram of Oriented Gradients (HOG), and Local Phase Quantization (LPQ) to capture texture information from facial regions. The features were then fed into an SVM for face PAD. Similarly, Bai et al. used the upper-body region and applied HOG to capture continuous edges of the presentation attack instrument (PAI). A linear SVM was utilized for detecting photo or video replay attacks [9], [12].

In 2013, Kose and Dugelay presented a static texture-based approach for 3D mask attack detection using LBP features [13]. Although 3D mask PADs were less studied due to the scarcity of public mask attack databases at that time, the LBP-based method effectively detected 3D mask attacks using the texture (original) image. They later enhanced this method by fusing LBP features from both texture and depth images, improving its detection capability [14]. In the same year, Galbally et al. introduced face PAD methods based on Image Quality Assessment (IQA), assuming that spoofing images captured in photos or video replays would exhibit different qualities than genuine samples [15], [16]. These IQA-based methods assessed image quality using various measures, such as sharpness, color and luminance levels, and structural distortions. The image quality scores were combined and fed into classifiers, such as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), for face presentation attack detection. The major advantage of the IQA-based methods was their non-reliance on priori face or body region detection, making them ''multi-biometric'' methods applicable to iris or fingerprint-based liveness detection. However, their performance was limited compared to other texture-based methods, and they were not designed to detect 3D mask attacks [15], [16].

In 2015, Boulkenafet et al. proposed extracting LBP features in HSV or YCbCr color spaces, leveraging subtle differences between genuine faces and presentation attacks [17], [18]. By simply changing the color space used, their LBP-based method achieved state-of-the-art performance compared to more complex methods based on Component Dependent Descriptor (CDD) [12] and even emerging deep Convolutional Neural Networks (CNNs) [19]. This work highlighted the significance of utilizing diverse color spaces for face PAD [17], [18].

B. DEEP LEARNING-BASED METHODS

Deep learning-based methods have also been explored to automatically learn texture features for face PAD. Deep learning techniques have found extensive application in various biometric authentication systems and face presentation attack detection [12] as well as in other various disciplines such as Natural Language Processing (NLP), autonomous driving and medical imaging and so on. These methods involve training deep networks using diverse patterns and leveraging the learned features to identify similar patterns within the dataset. Notably, deep learning excels in both supervised and unsupervised tasks, enabling efficient classification and clustering of data without the need for class labels. Yang et al. demonstrated the potential of using CNNs for PAD [19]. They applied a one-path AlexNet to learn texture features that effectively discriminate between genuine and presentation attack images. The method involved replacing the usual output of AlexNet with a Support Vector Machine (SVM) featuring binary classes. The approach outperformed existing methods when the input image was enlarged, highlighting the importance of context information from the background in face PAD. Following this breakthrough, more CNN-based methods were developed for face PAD [20], [21]. Patel et al. [20] proposed an end-to-end framework based on CaffeNet (a one-path AlexNet variant), utilizing a voting fusion strategy to achieve state-of-the-art performance for photo and video replay attack detection. Similarly, Li et al. utilized VGG-Face for face PAD, leveraging features from different layers of the CNN to improve performance. Their method demonstrated state-of-the-art results in both intra-dataset and cross-dataset scenarios [21]. Since then, numerous architectures have been devised for detecting photo and video replay attacks [22] [23].

In 2017, Arashloo et al. [24] introduced a novel approach to address the challenge of detecting unseen attacks in the context of anomaly detection. They formulated the problem

as a one-class classification task, considering real faces as the positive class and training a one-class SVM [25] to distinguish them. Similarly, in 2018, Nikisins et al. [26] employed one-class Gaussian Mixed Models (GMM) to model the distribution of genuine faces, enabling them to detect previously unseen attacks. Unlike the approach in [24], Nikisins et al. aggregated three publicly available datasets for training their model.

While the abovementioned methods solely used samples of genuine faces to train one-class classifiers, it is worth noting that known spoof attacks can also provide valuable insights for detecting novel and previously unseen attacks. To address this aspect, Liu et al. [27] proposed a CNN-based Deep Tree Network (DTN) in 2019, focusing on analyzing 13 attack types that encompass both impersonation and obfuscation attacks. Initially, the known presentation attacks (PAs) were clustered into eight semantic sub-groups using unsupervised tree learning, serving as the eight leaf nodes of the DTN. A Tree Routing Unit (TRU) was then learned to guide known presentation attacks to the appropriate tree leaf (i.e., sub-group) based on the features learned by the tree nodes (i.e., Convolutional Residual Unit (CRU)). Within each leaf node, a Supervised Feature Learning (SFL) module was employed, consisting of a binary classifier and a mask estimator, aiming to discriminate between different spoofing attacks. The mask estimation procedure is analogous to the depth map estimation, as previously presented by the same authors in [28]. Consequently, by utilizing the estimated mask and the score from a binary softmax classifier, unseen attacks can be effectively discriminated.

Jourabloo et al. proposed a GAN-based method, De-Spoof Net (DS Net), for estimating noise in spoof face images [29]. By assuming that the noise of a genuine image is zero, the method detected spoof images by thresholding the estimated noise. DS Net used different losses to model various noise patterns, achieving superior performance compared to other state-of-the-art deep face PAD methods [29]. Additionally, George et al. proposed Deep Pixelwise Binary Supervision (DeepPixBiS), based on DenseNet, which utilized pixel-wise binary cross-entropy loss along with binary cross-entropy loss for the final output. By forcing the network to learn patch-wise features, DeepPixBiS demonstrated promising performance for both photo and video replay attacks [30].

Image depth information plays a pivotal role in assessing the authenticity of a face, as real faces in the physical world possess three-dimensional structures, whereas faces captured in photographs or displayed on screens are two-dimensional. Even when the face presented in a photograph appears unnatural, the depth map exhibits discrepancies from that of a genuine face. Leveraging this insight, Atoum et al. pioneered the use of face depth maps for discriminating face spoofing attacks. In their study, a novel two-channel CNN-based face anti-spoofing method was proposed [31].

ResNet stands out as one of the most prevalent CNN architectures utilized in face anti-spoofing, capable of acquiring robust feature representations from facial images [22]. However, the manual design of neural networks remains a daunting task. Consequently, there has been a recent shift towards the automatic design of neural networks [32]. Regrettably, many of these methods rely on computationally expensive models, rendering them unsuitable for real-time Face Anti-Spoofing (FAS applications. In [33], Benlamoudi et al. presented an approach based on background subtraction and used pre-trained ResNet-50 [34] CNN architecture to learn features related to genuine and spoof faces. Their approach is based on the assumption that the capturing camera is always static.

Addressing the challenge of computational complexity has been a focal point in numerous studies, leading to the development of efficient and lightweight architectures for various computer vision tasks. One common approach involves the quantization of weights and/or activations of a base CNN model into lower-bit representations [35], or the pruning of unimportant filters based on floating-point operations per second (FLOPs) [36]. Alternatively, some methods entail the direct crafting of more efficient mobile architectures. For instance, MobileNet [37] heavily employs depthwise separable convolution to minimize computation density, while ShuffleNets [38] [39] utilize low-cost group convolution and channel shuffling. More recently, MobileNetV2 [40] has set a new benchmark for lightweight models in image classification by introducing inverted residuals and linear bottlenecks. Nevertheless, the design of hand-crafted models demands considerable human effort due to the potentially vast design space.

Several of the aforementioned lightweight mobile architectures have undergone modifications to enhance their discriminative and generalization capabilities for face recognition purposes [41], [42], and [43]. These specific models have been investigated across different face recognition scenarios, including image and video face recognition [44] and masked face recognition [45]. However, few works explore the potential of lightweight CNNs for video replay spoofing detection as a means of balancing accuracy performance with efficiency in real-world scenarios.

C. LITERATURE OVERVIEW AND THE RESEARCH GAP

As discussed in previous subsections, there are several approaches proposed in the literature for video-replay spoofing detection. Many of these methods rely on face detection as a preliminary step to localize facial information in the image before classification. However, it is crucial to acknowledge the drawbacks of face detection algorithms, as any failure in this step can lead to erroneous classification results. Moreover, conventional face detection algorithms are tailored for images containing only the faces of users, raising doubts about their adaptability to phygital game scenarios where images encompass diverse visual contexts with various foregrounds and backgrounds.

One significant visual phenomenon affecting phygital gaming images is the moiré effect, resulting from spatial

interference between the pixel grids of camera sensors and device screens [46]. The moiré effect is a visual phenomenon that occurs when two repetitive patterns or grids overlap, creating a new, often unexpected pattern. In the context of phygital gaming, the moiré effect typically arises from the interaction between the pixel grid of a camera sensor and the pixel grid of a device screen.

When a camera captures an image of a digital display, such as a smartphone or computer monitor, the pixel grid of the display may interfere with the pixel grid of the camera sensor. This interference can manifest as a moiré pattern, which appears as irregular, wavy lines or patterns superimposed on the image. These patterns can distort the original image and introduce visual artifacts. The moiré effect is highly sensitive to factors such as the relative alignment and spacing of the two grids, the resolution of the camera sensor and display, and the viewing angles [47]. Variations in these factors can lead to changes in the appearance of the moiré pattern, making it challenging to predict and mitigate.

Consequently, intricate moiré patterns manifest across different frequency bands of images with diverse colours, shapes, and intensities. Studies have also indicated that moiré pattern colours, especially brightness, exhibit slight discrepancies depending on the displayed background, rendering them non-robust [48], [49]. These factors emphasize the necessity for distinct approaches tailored specifically for spoof detection in phygital games, distinct from conventional methods employed in face biometric recognition.

One noteworthy endeavour in the literature towards addressing spoof detection in phygital games is by Huszár and Adhikarla [50]. They introduced a deep-learning-based approach capable of real-time parallel operation with HAR for presentation attack detection in videos. However, their approach also entails extracting facial information from images as an initial step for classification. In this paper, we present the first study applying deep-learning techniques for spoof detection in phygital games. Our primary contribution includes the introduction of a novel dataset comprising videos of users juggling footballs, supplemented by additional videos of replay attacks recorded on multiple monitors using various smartphone cameras. The design of our dataset is aimed at addressing the limitations of existing datasets for replay detection, particularly in phygital games, which are often constrained in terms of size and diversity. To the best of our knowledge, our dataset stands as the largest and most diverse collection for liveliness detection in phygital games.

To demonstrate the potential of our dataset, we develop a deep learning approach based on the EfficientNet [51] architecture for liveliness detection in phygital games. Our results showcase the effectiveness of the proposed approach, achieving high accuracy on the presented dataset. Additionally, we conduct an ablation study to evaluate the diversity and robustness of our dataset.

Our work significantly contributes to research endeavors aimed at enhancing the security and integrity of phygital games. The proposed dataset and deep learning approach serve as valuable resources for researchers and practitioners in developing more effective replay attack detection systems, thus fostering a safer phygital gaming environments.

III. OVERVIEW OF DATABASES FOR VIDEO-REPLAY ATTACK DETECTION

Currently, a notable research gap exists in the availability of public datasets specifically tailored for detecting video-replay attacks in the context of phygital games. However, datasets designed for face presentation attack detection in biometric recognition applications, such as face detection, are accessible and offer diverse attack scenarios, encompassing static 2D photo attacks, paper masking attacks, and 3D rigid and silicon mask attacks. Although these datasets may not directly align with the unique challenges of phygital games, where the focus is on assessing video genuineness rather than identifying individual players, they can still be valuable for training and validating deep learning models for video-replay attack detection in general. Some portions of the face presentation attack detection datasets that include video-replay attacks could be adapted and relevant for phygital gaming applications. In Table 1, we provide a comprehensive list of pertinent datasets containing video-replay attacks, sourced from the field of biometric face recognition.

It is evident from 1 that the variety of devices utilized for capturing and presenting videos to simulate replay attacks is notably scarce in the literature. Moreover, specific datasets, such as the one introduced by George et al. [59], feature cropped and resized videos focusing solely on facial information, which deviates from the patterns observed during actual replay video capture on a monitor. Consequently, models trained on such data may struggle to generalize to unseen scenarios. An effective deep-learning model for video-replay attack detection in phygital games necessitates a diverse set of samples recorded from multiple monitors and using various cameras. The lack of publicly available datasets addressing these requirements highlights a significant gap in developing efficient deep-learning approaches for video-replay attack detection.

To address the limitations in the literature and enable experimentation with video-replay attack detection in the context of phygital games, we have curated a new dataset comprising 600 participants engaged in football juggling sessions. During each session, participants performed simple football juggling tasks while being recorded by multiple cameras, including both IOS and Android smartphones. The dataset encompasses a diverse range of subjects, spanning different ethnicities (Asian, African, and Caucasian), ages (children and adults), and genders (male and female). The recording sessions were conducted under various environmental and lighting conditions, encompassing both indoor and outdoor settings, as well as different times of day, including daytime and nighttime conditions.

Moiré effects, that occur while capturing the video replay on a device screen using any given camera, have been

TABLE 1. A compilation of publicly available datasets in the literature that include examples of video-replay attacks. For the samples marked with an asterisk, the reported number of samples encompasses print and 2D photo attacks, and the number of samples containing video-replay attacks is notably smaller than the total count.

Dataset	# Subjects	#Samples	Original videos resolution	# Devices used (Capture/Replay)
Replay-Attack (2012) [52]	50	650	Varying resolution	1/2
CASIA-FASD (2012) [53]	50	200	Varying resolution	3/1
MSU-MFSD (2015) [54]	35	280	720 × 480, 640 × 480	2/2
MSU-USSA (2016) [55]	1140	**10260***	1920 × 1080	2/3
Replay-Mobile (2016) [56]	40	1300*	Varying resolution	2/2
OULU-NPU (2017) [57]	55	2970	Varying resolution	6/2
SiW (2018) [28]	165	4478*	1920 × 1080	1/4
ROSE-Youtu (2018) [58]	20	3350	640 × 480, 1280 × 720	5/2
WMCA (2019) [59]	72	695	1920 × 1080	2/1
CASIASURF CeFA (2020) [60]	**1607**	7200*	1280 × 720	-
VFPAD (2022) [61]	40	5836*	1280 × 720(12-bit)	1/1
Sqiller-Spoof (ours)	600	2700	1280 × 720	**7/17**

FIGURE 3. Illustration of Video-Replay Spoofing Scenarios in a phygital football game. The left side demonstrates a video replay attack when the camera is positioned close to the device screen, resulting in visible Moiré interference patterns. On the right side, the camera moves away from the device screen such that the device screen bezels are visible, causing reduced severity of the Moiré interference patterns. These scenarios showcase the challenges in spoof detection in the presence of visible device screen bezels.

recognized as highly sensitive to various factors, particularly variations in camera pose and the distance between the camera and the device screen [46]. In our context of phygital games, such effects can significantly impact the integrity of liveliness detection. Figure 3 illustrates this phenomenon with two cases of video replay spoofing in Sqiller, a phygital gaming environment centred around football juggling.

Sqiller's functionality relies on detecting and tracking the trajectory of the football and the body coordinates of the athlete to assess their performance. When malicious users attempt to spoof Sqiller, they often aim to mimic the real gaming environment by staying in proximity to the device screen, ensuring the visibility of both the player and the football. As a result, the camera's distance from the screen plays a crucial role in determining the severity of Moiré interference patterns, as depicted in Figure 3.

In the left side of Figure 3, when the camera is close to the device screen, Moiré interference patterns become more pronounced. Conversely, on the right side, when the camera is moved further away from the screen, the severity of Moiré patterns diminishes (the extent to which the severity diminishes depends on the device screen resolution). However, gaming logic establishes an upper limit on how far the camera can be positioned from the device screen to maintain effective gameplay. Our observations with Sqiller indicate that gaming logic still functions when the device screen bezels are visible, as shown on the right side of Figure 3. This poses a challenging problem for spoof detection, as the frequencies of interference patterns may not always align with realizable scenarios. Moreover, detecting spoofing attempts when the device screen bezels are visible has received limited attention in face biometric spoof detection literature [62]. Traditional cross-entropy loss may not be effective in training deep learning models for such scenarios, as networks can inadvertently learn arbitrary patterns, such as screen bezels, instead of essential spoof patterns [28]. This can lead to reduced detection accuracy while detecting video replay spoofing. To address this issue, we dedicate a new class of videos that also contain video-replay spoofs with device screen bezels visible.

To this end, we introduce our novel dataset comprising 900 genuine videos of 600 users engaging in football juggling tasks, accompanied by 900 video-replay attack videos where no device screen bezels are visible, and an additional 900 video-replay attack videos where device screen bezels are visible. The replay dataset is created using 7 distinct capture devices and 17 replay devices in 18 sessions, with each session containing 50 videos. To maintain dataset balance, each genuine video is uniquely employed when generating replay-attack videos. The creation process involves displaying the videos on diverse monitors, encompassing smartphone screens to wide-screen televisions while recording the resulting videos through various IOS and Android cameras. On average, the captured videos have a duration of approximately 3 seconds. A visual depiction of select samples from the dataset is provided in Figure 4.

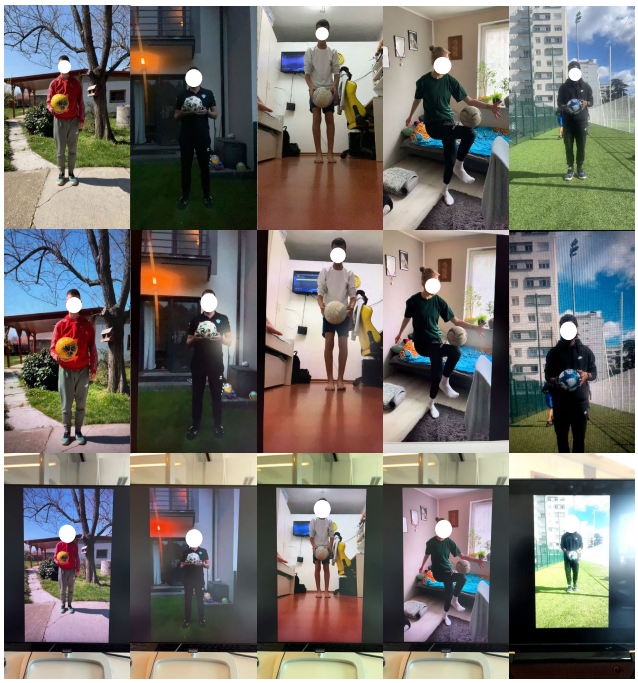

FIGURE 4. Sample frames from our Sqiller-Spoof dataset for video-replay attack detection in phygital games. The top row displays genuine videos, the middle row displays simulated video-replay attacks captured using various monitors and cameras and the bottom row displays simulated video-replay attacks when the device screen bezels are visible. The faces of users have been masked to preserve their privacy.

The introduction of this comprehensive dataset is anticipated to substantially advance the field of deep-learning models designed for discerning video-replay attacks within phygital games environments.

IV. VIDEO-REPLAY SPOOFING DETECTION IN PHYGITAL GAMES

In recent years, deep Convolutional Neural Networks (CNNs) have emerged as a promising approach for detecting video-replay attacks, surpassing traditional statistical methods in performance [63], [64]. However, a common challenge faced in training CNNs is the scarcity of available data, which can hinder their training from scratch. To overcome this limitation, most CNN-based video-replay attack detection methods adopt a two-step approach. Initially, a CNN is trained end-to-end on image recognition tasks using publicly available datasets like ImageNet [65]. Subsequently, the CNN is fine-tuned using specialized video-replay attack datasets to discern between genuine and attack presentations. This fine-tuning process involves leveraging the outputs of intermediate layers from the pre-trained CNN and constructing additional layers and a classifier to enhance detection performance.

For our video replay attack detection in phygital games, we employed the pre-trained EfficientNet [51] architecture. The selection of EfficientNet was guided by several factors. Firstly, EfficientNet is a well-designed architecture that strikes a balance between model size and performance, making it particularly suitable for real-time applications. By employing depth, width, and resolution scaling, EfficientNet achieves state-of-the-art performance while utilizing fewer parameters compared to popular alternatives such as ResNet [34] or VGGNet [66]. This is especially advantageous for replay-attack detection in phygital games, as the model can efficiently run in real-time even on low-powered devices. Secondly, EfficientNet has demonstrated impressive results on various computer vision tasks, including image classification, object detection, and segmentation. This suggests that the architecture may effectively capture relevant features and patterns present in our juggling football videos, contributing to successful video-replay attack detection. Additionally, as EfficientNet has been trained on a vast dataset like ImageNet, it has acquired rich and generalizable features that can be fine-tuned to our specific problem of video replay attack detection.

In this study, we adapted the EfficientNet-B0 architecture, originally designed for image classification on ImageNet, for video-replay attack detection. To accomplish this, we tailored the final classification head to produce three logits corresponding to three specific classes. The system architecture is illustrated in Figure 5.

During the training phase, we assigned ground truth labels of 0 to genuine samples, 1 to replay attack samples, and 2 to replay attack samples with visible display bezels. Our system closely follows the design of the EfficientNet-B0 network, employing an input image resolution of 224×224. To construct the training dataset, we pre-processed the training videos to extract 32 video frames, uniformly sampled in the temporal domain, ensuring consistent representation regardless of video duration. From each frame obtained from genuine and spoofed videos, we extracted 15 non-overlapping cropped patches, each sized 224×224, for training purposes. For temporal frames extracted from spoofed videos containing display bezels, we uniformly sampled 15 patches in the spatial domain, originally of resolution 720×720, which were then resized to the target resolution of 224×224. This selection of large-sized patches aimed to capture global contextual information relevant to the presence and characteristics of bezels.

Our curated training dataset comprises a substantial number of approximately 1.3 million samples, covering the instances from all three classes. Prior to commencing each training session, we transformed the pixel value range of the extracted frames to fall within the $[0, 1]$ interval, yielding floating-point images. Furthermore, we normalized the video frames using mean and standard deviation to ensure consistency in the training process.

A. LEARNING AND OPTIMIZATION

During the learning phase, we applied data augmentation techniques to our training samples, including random horizontal flipping, Gaussian blurring, and random sharpness

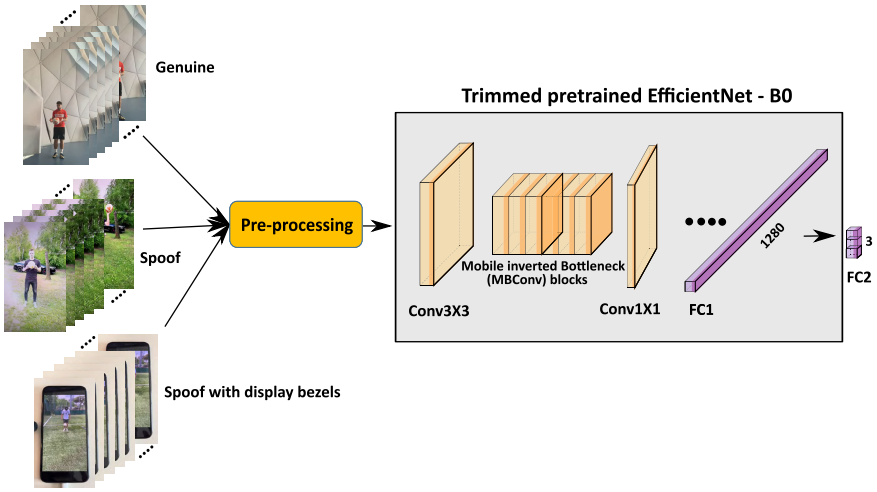

FIGURE 5. The system architecture illustrating the adapted EfficientNet-B0 model for video-replay attack detection. Specific alterations to the final classification head ensure the generation of three logits corresponding to the classes: 'Genuine,' 'Replay Attack,' and 'Replay Attack with Display Bezels Visible'.

TABLE 2. In our system, we have 4011391 parameters that are involved in the combination of the trimmed EfficientNet-B0 model and the replaced final classification head. Since the parameters of the trimmed EfficientNet-B0 model are also optimized during training, all 4011391 parameters are trainable.

Component	Output Shape	Params #
Trimmed EfficientNet-B0 model	(1280)	4007548
Dense ("FC2" in Fig. 5)	(3)	3843
Total Parameters	4011391	

TABLE 3. List of hyperparameters and their corresponding values used in our trainings.

Hyperparameter	Value
Learning Rate	0.0001
Batch Size	48
Number of Epochs	10
Optimizer	Adam
Loss Function	Cross Entropy loss

adjustment. These augmentations were incorporated to achieve a more comprehensive representation of our training dataset. We performed weight updates on all layers during training, including the feature extraction layers. The optimization process employed the Adam optimizer with a learning rate of 0.0001, while the Cross-Entropy loss function was used to guide the learning process.

The training procedure spanned 10 epochs, during which the modified network encompassed approximately 4.0 million parameters (as indicated in Table 2). This parameter count facilitated the effective updating of all network parameters through back-propagation, enhancing the training process. For convenience and easy reference, all hyperparameters employed in our study are listed in Table 3.

To ensure consistency, within each training batch, we concatenated training data from all three classes, resulting in an effective batch size of 48 samples. This approach enabled efficient utilization of available training data during the optimization process.

V. RESULTS AND DISCUSSION

A. EXPERIMENTING WITH OUR SQILLER-SPOOF DATABASE

To ensure a robust evaluation of our proposed data set using our deep learning-based approach, we performed a rigorous 5-fold cross-validation study. This cross-validation approach guarantees an extensive assessment of the model and mitigates any potential overfitting issues arising from specific training and testing data splits. Specifically, we partitioned the data set into 5 systematic and non-overlapping splits, each containing 80% of samples for training and 20% for testing, for comprehensive evaluation. Our findings from the 5-fold cross-validation study are presented in Table 4 utilizing a selection of metrics frequently employed in the assessment of biometric presentation attacks: the Attack Presentation Classification Error Rate (APCER), Bona fide Presentation Classification Error Rate (BPCER), and the Half-Total Error Rate (HTER) as outlined in ISO standards [67]. Analogous to the False Acceptance Rate (FAR) and False Rejection Rate (FRR), APCER and BPCER quantify the model's effectiveness. Hence, diminished values of APCER, BPCER, and subsequently HTER signify enhanced performance of the model. Additionally, we present the performance assessment using accuracy [68], rated on a scale from 0 to 1, where a value closer to 1 indicates superior performance. The results from our cross-validation studies highlight the consistency of our dataset and the effectiveness of our approach leveraging the EfficientNet B0 architecture for video replay attack detection.

To further examine the feature set learned by our model using the sqiller-spoof database, we visualized the learned features of our model utilizing t-SNE [69], dimensionality reduction technique. Specifically, we used the trained models from five cross-validation folds and performed

TABLE 4. The table presents the evaluation results of the modified EfficientNet-B0 architecture on the presented dataset using a five-fold cross-validation study. The evaluation is carried out using various popular quality metrics in the field of video replay attack detection. The results in the table demonstrate the effectiveness of the proposed method for detecting video replay attacks in the evaluated dataset.

Fold	Accuracy (attack, attack with visible bezels, genuine)	APCER	BPCER	HTER
1	96.46 (94.03,99.99,95.35)	2.99	4.65	3.82
2	96.59 (95.74,99.78,94.25)	2.24	5.75	3.99
3	96.28 (95.07,99.06,94.71)	2.94	5.29	4.11
4	96.98 (95.73,99.64,95.56)	2.31	4.44	3.38
5	96.16 (93.48,98.29,96.69)	4.11	3.31	3.71

feature visualization using the testing set. To accomplish this, we collected features before the last fully connected layer of the trained model using the testing dataset. These features represent the high-dimensional feature space of the detections in the testing set. In other words, such visualization provides a visual representation of the distribution of the testing set in the feature space according to the learned patterns from the training set. Figure 6 shows this feature visualization for a random set containing 250 test samples for each class in all the five cross-validation folds in order from left to right. In Figure 6, the colours of the points represent the corresponding labels of the images - 0 refers to genuine samples (blue), 1 refers to replay attack samples (green), and 2 refers to replay attack samples with visible display bezels (red). The t-SNE algorithm was used to reduce the dimensionality of the data, resulting in two dimensions that were constructed in a way that preserves the pairwise similarity of the high-dimensional features as much as possible. The resulting scatter plot allows for the identification and understanding of patterns in the data, as well as the potential identification of outliers where a trained model misclassifies any test samples.

Our results showed a natural clustering of the points, indicating that the model was able to effectively capture the relevant features and patterns in the data. However, it is also noticeable that there are potential outliers where the trained model misclassified the samples in all five folds. The observed outcomes in Figure 6 illustrate a significant degree of class overlap between class 1 and class 2. This outcome was anticipated given the commonality of replay-attack samples within these classes. However, it is worth noting that class 2 samples are captured at a slight distance from the device display, revealing the bezels that attenuate the strength of the Moiré effect. Furthermore, the discernible disparity between classes 0 and 1 underscores the effectiveness of our dataset in encompassing distinct samples that accurately differentiate genuine instances from video-replay attacks.

B. BACKBONE ADAPTATION: LEVERAGING ESTABLISHED ARCHITECTURES FOR REPLAY ATTACK DETECTION IN FACE RECOGNITION SYSTEMS

Efforts have been made in the past, as documented in the related works section, to address a similar issue: video-replay spoof detection in face biometric recognition. Many successful methods, rooted in deep learning, have adopted various lightweight architectures [70], such as MobileNetV2 [40] and ShuffleNetv2 [39], for both training and inference tasks. To assess whether retraining the face detection replay attack detection framework on the dataset proposed in this work can yield similar performance, we also trained these lightweight architectures on our dataset. Specifically, we utilized pre-trained models of these lightweight architectures trained on ImageNet [65] to ensure consistency in our comparisons with EfficientNet-B0, which is also trained on the same dataset. Our findings are presented in Table 6.

TABLE 5. Details of training and testing datasets for architecture implementation.

Class	Number of Videos	Percentage for Training	Percentage for Testing
Genuine	900	80%	20%
Replay Attack	900	80%	20%
Replay Attack with Display Bezels Visible	900	80%	20%

For this and subsequent experiments, we use the data split from our first cross-validation fold. Specifically, we consider 80% of samples for training and 20% for testing. We have selected 80% of samples from each of the 18 sessions (data from 40 videos in a session) for training, and 20% of samples from each session (data from 10 videos in a session) for testing. Table 5 provides details of the training and testing datasets used for architecture implementation.

According to our testing results, our approach employing the Efficient-B0 backbone achieved approximately 1% higher accuracy than the aforementioned lightweight architectures. It is noteworthy that a 1% improvement may seem insignificant, especially considering that these lightweight architectures have relatively fewer model parameters, which may not always lead to faster inference speeds but certainly require fewer computing resources. Hence, our results suggest that in resource-constrained scenarios, lightweight architectures, such as those considered in this experiment, can be suitable for efficient deployment. It's important to highlight that EfficientNet-B0, while efficient, does not impose significantly higher computational demands compared to larger architectures such as ResNet50 [34]. Therefore, when deploying an architecture for video-replay spoof detection in phygital games, careful consideration of system requirements is crucial. Notably, our experimental results on our dataset align with the observations made by Martínez-Díaz et al. [70], indicating that lighter-weight architectures do not result in a higher drop in accuracy.

Given that our dataset is pioneering in the field of video-replay spoofing detection within phygital games, a direct comparison of our model's performance with other deep learning-based methods is infeasible. To investigate the effectiveness of our approach, we turn to recent baseline techniques for spoofing detection, as the current literature does not encompass data-driven methodologies for detecting

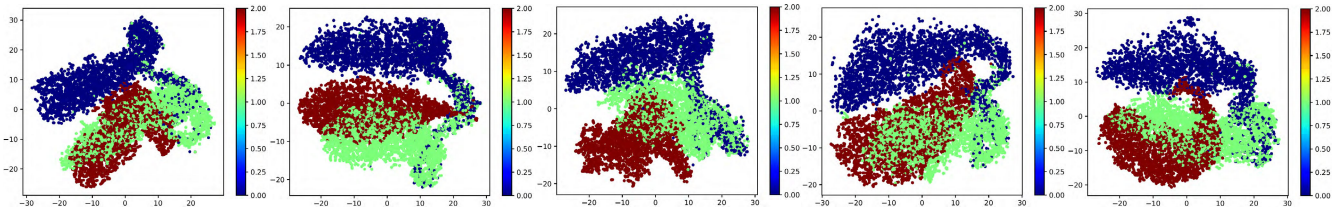

FIGURE 6. t-SNE visualization of the learned features from a video replay attack detection model. Each point represents a testing sample and the colours indicate the corresponding label (genuine, replay attack and replay attack with visible display bezels). The plot shows the natural clustering of the points that can be used to identify and understand patterns in the data.

TABLE 6. The table presents the results of evaluating our proposed approach against other light-weight and baseline spoofing detection methods using APCER, BPCER, and HTER metrics on the proposed dataset for video-replay spoof detection in phygital games. The lower values of these metrics indicate superior performance.

Backbone	Accuracy (attack, attack with visible bezels, genuine)	APCER	BPCER	HTER
LBP-SVM [71]	71.25 (77.55,86.32,49.87)	18.06	50.13	34.1
HOG-SGD [72]	46.80 (54.21,41.22,44.98)	52.01	54.78	53.39
MobileNetV2 [40]	95.98 (96.36,98.87,92.72)	2.39	7.28	4.83
ShuffleNetv2 [39]	95.44 (94.23,99.85,92.23)	2.96	7.77	5.36
EfficientNet-B0 [51]	96.46 (94.03,99.99,95.35)	2.99	4.65	3.82

video-replay attacks where users have the autonomy to select their positions during physical activities.

In this experimentation, we further investigate two descriptors - LBP histograms and HOG features - both of which have been documented as prospective feature descriptors for spoofing detection using facial image data (as referenced in the related works section). The outcomes of this comparative study are showcased in Table 6, affirming our approach's superiority over the baseline alternatives in spoofing detection. For the comparison involving LBP histograms, we employ an SVM trained through the scikit-learn toolkit, with the descriptor calculations adhering to a radius of $R = 1$ and a number of points $P = 8$, as elaborated in [73]. For the comparison involving HOG descriptors, we initiate by extracting HOG descriptors from the training images using the following parameters: orientations - 9, pixels per cell - (8, 8), and cells per block - (2, 2). Given memory constraints, here, we adopt a stochastic gradient descent (SGD) learning strategy with multiple batches.

The comparison is conducted on our first cross-validation fold, utilizing both training and testing samples. Table 6 illustrates the outcomes of the baseline techniques from the literature alongside our approach based on the EfficientNet-B0 architecture. The experimental results distinctly underscore our approach's substantial performance gain over the baseline methods. Our findings align with the conclusions drawn in [74], accentuating that strategies employing low or medium-level texture descriptors coupled with simple classifiers for spoofing detection do not fare well under cross-dataset evaluation protocols.

C. EXPLORING GENERALIZATION: EXPERIMENTS WITH FACE BIOMETRIC SPOOFING DATASETS TO ASSESS MODEL PERFORMANCE

In order to assess the generalizability of our model, which has been trained on the sqiller-spoof dataset, across diverse domains of video-based spoofing detection, we conducted a series of experiments utilizing various publicly available datasets commonly employed for spoof detection within facial recognition systems. These datasets include ROSE-Youtu [58], CASIA-FASD [53], Replay-Attack [52], Replay-Mobile [56], SiW [28], SiW-Mv2 [27], and OULU-NPU [57].

Our experimentation methodology involved applying the model trained on the sqiller-spoof dataset to these distinct datasets for the purpose of generalization analysis. Notably, it is highlighted by Yang et al. [48] that Moiré patterns lack resilience when confronted with background variations and furthermore, visual dissimilarities exist between perceptually varying settings such as phygital games and facial biometric recognition contexts. The majority of facial biometric spoof datasets encompass not only video replay attacks, but also encompass a range of other attack modalities, including 2D photo attacks, paper masking attacks, and 3D rigid silicone mask attacks. For each aforementioned dataset, we specifically focused on the designated testing subset as provided by the original dataset authors. Among the testing video samples, we only consider those instances associated with video replay-based spoofing attacks. By exclusively focusing on video replay attack instances within the chosen datasets, it's essential to recognize that our results cannot be readily compared to methodologies encompassing broader attack types. Traditional approaches often incorporate a wider array of attack modalities in their training and validation processes, necessitating careful consideration of this discrepancy when interpreting our experimental findings.

1) VIDEO LEVEL PREDICTION

Our trained efficientNet-b0 model generates image-level predictions. To extend these predictions to the video clip level, we devised a strategy for aggregating frame-level predictions within a video, enabling the classification of the video as either genuine or an attack. Notably, the publicly

available datasets for spoofing detection in face biometric recognition applications lack the explicit class "replay attack with visible bezels." Within the samples of replay attacks, videos occasionally feature visible display bezels.

For video-level inference, we adopt a two-step approach. Initially, we assess the presence of display bezels across multiple frames. If these bezels are detected in the majority of considered frames, the video is promptly classified as containing a video replay attack. In the absence of detected bezels, we proceed to evaluate the presence of learned moiré patterns across multiple frames. Should these patterns manifest in over 50% of frames, we classify the video as an attack; otherwise, it's deemed genuine.

Our process begins with the extraction of 32 candidate frames from a given video, uniformly sampled across the temporal domain. Through cropping and/or resizing, we prepare a batch of 32 images at a resolution of 224 × 224 as input to the model, resulting in 32 predictions (refer to Figure 7). For bezel detection, predictions from the initial 8 frames are used; if more than 4 frames are identified with bezels, the video is classified as an attack.

To accomplish this, we extract the 8 largest square-shaped patches possible from the first 8 frames. These patches adapt to the frame's dimensions (height x height if width < height; otherwise, width x width). This process, shown in the upper right of Figure 7, utilizes a sliding window approach to ensure comprehensive spatial coverage. These substantial patches encapsulate global contextual cues, facilitating bezel detection.

Subsequently, we employ predictions from the remaining 24 frames to detect the presence of Moiré patterns. To achieve this, the central regions of these frames are cropped into 224 × 224 segments (as depicted in the bottom three rows on the right side of Figure 7), yielding 24 local predictions. A video is categorized as an attack if more than 12 of these local predictions indicate the presence of Moiré patterns; otherwise, it's classified as genuine. The pseudo-code for the described method is provided in Algorithm 1.

We report the generalization ability of our model using the same performance metrics as in cross-validation experiments. Table 7 presents the performance evaluation of our model across the considered datasets. For each dataset, we furnish details encompassing the testing set's sample count, pixel-based video resolutions embedded within each dataset, and the temporal duration of the videos. The outcomes indicate commendable model performance on the majority of face biometric recognition datasets, notably excelling in cases where the video resolution closely aligns with that of our training videos.

A clear trend appears in the results: the Replay-Mobile dataset [56] exclusively features videos at a resolution of 720 × 1280 pixels, closely resembling our training video resolution. As anticipated, our model attains peak performance on this dataset. ROSE-Youtu [58], CASIA-FASD [53], and SiW-Mv2 [27] datasets also encompass a subset of videos at 720 × 1280 resolution, where our model achieves a substantial accuracy of approximately 0.9. Conversely, the Replay-Attack dataset [52], characterized by 320 × 240 resolution videos, falls below our training video resolution, leading to sub-optimal generalization and performance.

Algorithm 1 Video Analysis and Prediction

```
Input  : video
Output: Result
Initialize indices ← Get32UniformTemporalIndices(video);
Initialize bezelCount ← 0;
for i ← 1 to 8 do
    frame ← ExtractFrame(video, indices(i));
    prediction ← ModelPredict(frame);
    if prediction indicates presence of bezels then
        bezelCount ← bezelCount + 1;
if bezelCount > 4 then
    return "Attack";
Initialize moireCount ← 0;
for i ← 9 to 32 do
    frame ← ExtractFrame(video, indices(i));
    prediction ← ModelPredict(frame);
    if prediction indicates presence of Moiré patterns then
        moireCount ← moireCount + 1;
if moireCount > 12 then
    return "Attack";
else
    return "Genuine";
```

The OULU-NPU dataset [57] contains videos in full HD, and our model's generalization to this dataset remains unsatisfactory. On the other hand, the SiW dataset [28], also with full HD videos, shows better model performance. The difference lies in visible display bezels in SiW's testing videos, which prevents our method from progressing to Moiré pattern detection.

These results align with Yang et al.'s observations [48], emphasizing Moiré pattern sensitivity to scales and resolutions. Our generalization study outcomes reinforce the value of our dataset, affirming its capacity to effectively train our model in acquiring meaningful detection features. Our findings also proove that our model has genuinely acquired relevant attributes instead of random characteristics, thus strengthening the validity of our approach. Furthermore, the classification error rates, APCER and BPCER provide another insightful observation in our results. It is evident that, in most instances, our model consistently demonstrates lower error rates in predicting genuine cases, as evidenced by BPCER scores being notably lower than APCER scores. This consistency holds across various video resolutions. This phenomenon is of paramount significance due to the inherent risk associated with misclassifying a genuine video as a spoof video. Ensuring that such misclassifications are minimized is crucial to upholding the integrity of applications like phygital games, where accurate classification holds substantial implications.

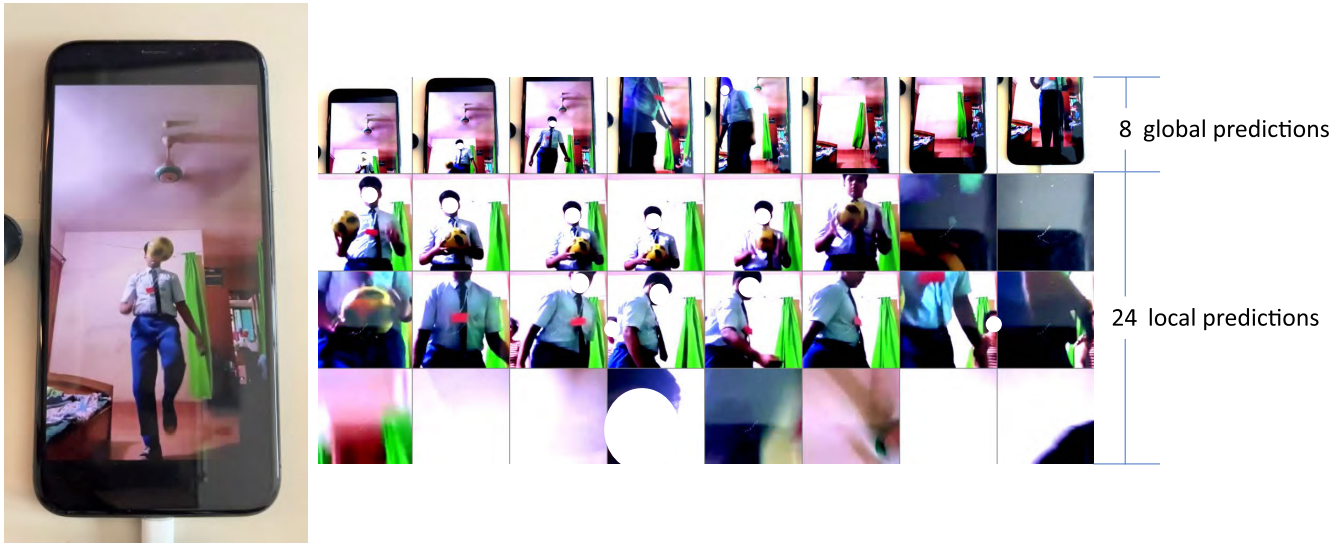

FIGURE 7. Thirty-two frames are selected for video-level prediction. The left side displays a video frame extracted from a sample testing video within our dataset, while the right side showcases the 32 extracted patches from these sampled frames. These patches encompass both global and local cues, contributing to precise video-level predictions.

TABLE 7. Performance of our model on face biometric recognition spoofing datasets using various metrics. The table also presents sample counts, video resolutions, and duration for each dataset, providing insights into our model's effectiveness across various scenarios.

Dataset	Accuracy (attack, genuine)	APCER	BPCER	HTER	Resolution	Samples (Attack ; Genuine)	Avg. Duration (min ; max) in sec.
ROSE-Youtu [58]	90.46(82.25, 98.66)	17.75	1.34	9.54	480X640 ; 720X1280	849 (400;449)	11.52 (6.71;19.6)
CASIA-FASD [53]	85.8(71.59, 100)	28.41	0.0	14.2	480X640 ; 640X480 ; 720X1280	176 (88;88)	7.07 (1.92;19.52)
Replay-Attack [52]	63.12(26.25, 100)	73.75	0.0	36.88	320X240	240 (160;80)	11.26 (9.2;15)
Replay-Mobile [56]	95.38(91.67, 99.09)	8.33	0.91	4.62	720X1280	158 (48;110)	9.9 (8.9;10)
SiW [28]	87.93(75.86, 100)	24.14	0.0	12.07	1080X1920; 1920X1080	312 (203;109)	17.17 (0.03;32.58)
SiW-Mv2 [27]	76.54(53.85, 99.23)	46.15	0.77	23.46	720X1080; 1920X1080	300 (261;39)	5.9 (1.2;10)
OULU-NPU [57]	63.89(27.78, 100)	72.22	0.0	36.11	1080X1920	1080 (720;360)	4.96 (3.09;15.06)

TABLE 8. Accuracy (%) comparison of lightweight architectures (MobileNetV2, ShuffleNet V2) and EfficientNet-B0 on publicly available datasets for face biometric spoof detection..

Dataset	% Accuracy (attack, genuine)		
	ShuffleNetv2 [39]	MobileNetV2 [40]	EfficientNet-B0 [51]
ROSE-Youtu [58]	81.8(64.5, 99.11)	89.62(81.25, 98.00)	**90.46**(82.25, 98.66)
CASIA-FASD [53]	74.43(52.27, 96.59)	63.07(26.14, 100.00)	**85.80**(71.59, 100.00)
Replay-Attack [52]	57.5(17.5, 97.5)	58.75(17.50, 100.00)	**63.12**(26.25, 100.00)
Replay-Mobile [56]	91.67(83.3, 100.0)	89.13(79.17, 99.09)	**95.38**(91.67, 99.09)
SiW [28]	83.74(67.49, 100.0)	81.77(63.55, 100.00)	**87.93**(75.86, 100.00)
SiW-Mv2 [27]	72.69(46.15, 99.23)	69.94(41.03, 98.85)	**76.54**(53.85, 99.23)
OULU-NPU [57]	62.71(29.03, 96.39)	62.57(27.64, 97.50)	**63.89**(27.78, 100.00)

D. ABLATION STUDY

We also conducted ablation studies to assess the generalization capabilities of the lightweight architectures, MobileNetV2 and ShuffleNet V2, to unseen attacks. Specifically, we trained these lightweight architectures on our training dataset and evaluated their performance on publicly available datasets for face biometric spoof detection, as outlined in Table 7. The results of this ablation study, measured in terms of accuracy (%) and Half Total Error Rate (%HTER), are presented in Tables 8 and 9 respectively.

Our findings indicate that our approach utilizing the EfficientNet-B0 backbone consistently outperformed the lightweight architectures. Contrary to the observations made by Martínez-Díaz et al. [70], our results demonstrate that, for most datasets, our approach significantly surpasses the performance of considered lightweight architectures. Specif-

TABLE 9. Half Total Error Rate (%HTER) comparison of lightweight architectures (MobileNetV2, ShuffleNet V2) and EfficientNet-B0 on publicly available datasets for face biometric spoof detection..

Dataset	% HTER (APCER, BPCER)		
	ShuffleNetv2 [39]	MobileNetV2 [40]	EfficientNet-B0 [51]
ROSE-Youtu [58]	18.20(35.5, 0.89)	10.38(18.75, 2.00)	**9.54**(17.75, 1.34)
CASIA-FASD [53]	25.57(47.73, 3.41)	36.93(73.86, 0.00)	**14.20**(28.41, 0.00)
Replay-Attack [52]	42.5(82.5, 2.5)	41.25(82.50, 0.00)	**36.88**(73.75, 0.00)
Replay-Mobile [56]	8.33(16.67, 0.0)	10.87(20.83, 0.91)	**4.62**(8.33, 0.91)
SiW [28]	16.26(32.51, 0.0)	18.23(36.45, 0.00)	**12.07**(24.14, 0.00)
SiW-Mv2 [27]	27.31(53.85, 0.77)	30.06(58.97, 1.15)	**23.46**(46.15, 0.77)
OULU-NPU [57]	37.29(70.97, 3.61)	37.43(72.36, 2.50)	**36.11**(72.22, 0.00)

TABLE 10. Performance comparison of architectures and baseline methods on the dataset proposed by Huszár and Adhikarla [50] within the phygital sports domain, using metrics including APCER, BPCER, and HTER.

Method	APCER	BPCER	HTER
LBP [75]	40.5780	48.1586	44.3683
LBP-TOP [76]	27.2727	40.3226	33.7977
SBP [75]	3.1915	45.5556	24.3735
SBP-TOP [73]	15.4124	78.9660	47.1892
EM [50]	8.9109	6.1151	7.5130
ShuffleNetv2 [39]	0	12.5	6.25
MobileNetV2 [40]	0	0	**0**
EfficientNet-B0 [51]	12.5	0	6.25

ically, our results suggest that lightweight architectures, particularly those examined in our experiment, do not generalize well to unseen attacks. Conversely, our approach based on the EfficientNet-B0 architecture exhibits robust generalization to various unseen scenarios. Here, we emphasize that Martínez-Díaz et al. [70] conducted their study across multiple spoof attacks, including print, paper, replay, and 3D mask attacks. In contrast, our experiments specifically focused on video-replay attacks. This distinction in the nature of attacks considered underscores the importance of tailoring detection methodologies to the specific threat landscape encountered in different contexts.

For further exploration of the performance of the considered architectures in our experiment, we also computed and presented confusion matrices and Precision-Recall curves for various architectures across different datasets used for generalization experiments. The confusion matrices for ShuffleNet V2, MobileNetV2, and EfficientNet-B0 across various datasets are depicted in Figures 8, 9, and 10 respectively.

From the confusion matrices, it is evident that our trained model based on EfficientNet-B0 produced fewer false negatives (genuine videos classified as attacks) and false positives (attack videos classified as genuine).

The Precision-Recall curves for ShuffleNet V2, MobileNetV2, and EfficientNet-B0 are presented in Figures 11, 12, and 13 respectively, illustrating the trade-off between precision and recall at different classification thresholds for various datasets. A higher Precision-Recall curve signifies better classifier performance. Each Precision-Recall curve plot also includes the Area Under Curve (AUC) value. A high AUC value indicates that the classifier effectively identifies genuine videos (high precision) and captures as many of them as possible (high recall), which is crucial for accurately detecting video-replay spoof attacks.

The AUC values demonstrate that akin to accuracy percentage, our approach based on EfficientNet-B0 consistently exhibited superior performance across various public datasets for face biometric spoof detection.

We conducted further investigation into the performance of these architectures using a dataset proposed by Huszár and Adhikarla [50] within the same phygital sports domain. The results are detailed in Table 10. Additionally, we present the performance of other baseline methods previously introduced for comparison on this dataset. Following the metrics employed in the study by Huszár and Adhikarla [50], we present the results using the metrics: APCER, BPCER, and HTER.

Our experiments with this dataset reveal that MobileNet V2 demonstrated superior generalization abilities compared to other architectures and baseline methods utilized in the comparison. It's important to note that we provide these results for the sake of completeness. However, it's crucial to emphasize that there are only 7 videos available for testing in this dataset. Such a small testing set may not adequately represent the diversity and complexity of real-world data that a model encounters. Consequently, the performance metrics obtained on the testing set may not accurately reflect how well the models generalize to unseen data.

With a small testing set, there is diminished statistical confidence in the performance metrics derived from evaluating the models. This can impede the ability to draw meaningful conclusions about the effectiveness of these model and the significance of any observed differences in performance between different models or parameter settings.

E. IMPLEMENTATION AND PERFORMANCE

We developed a lightweight, independent spoof detection system tailored for phygital games by utilising the NVIDIA DeepStream SDK [77]. Our model was constructed using the pytorch deep learning library [78], involving training on our dataset. Subsequently, the trained model underwent conversion into a suitable format that facilitates seamless integration and execution within the DeepStream environment. In this self-contained framework, video-level predictions followed the procedure outlined in Algorithm 1. Specifically, 32 image

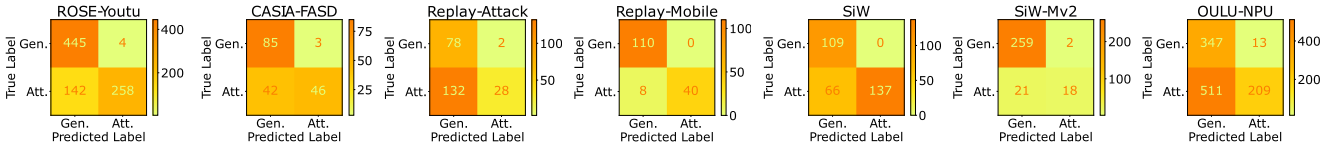

FIGURE 8. Confusion matrix illustrating the performance of ShuffleNet V2 across various datasets used for generalization experiments. The matrix provides a visual representation of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications, indicating the model's ability to distinguish between genuine and attack videos.

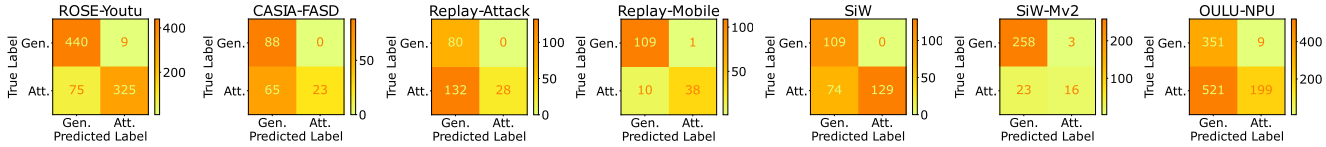

FIGURE 9. Confusion matrix displaying the performance of MobileNetV2 across different datasets used for generalization experiments. The matrix delineates the model's classification accuracy in terms of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances, shedding light on its efficacy in identifying genuine and attack videos.

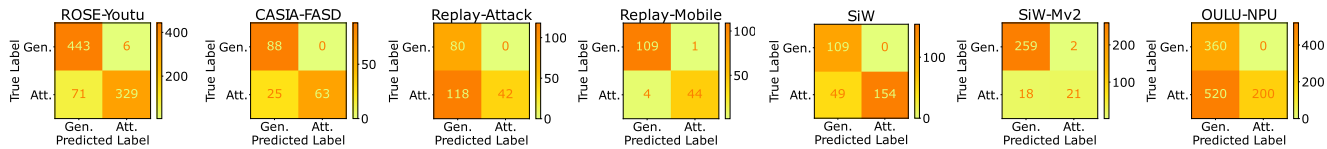

FIGURE 10. Confusion matrix demonstrating the performance of EfficientNet-B0 across a variety of datasets utilized for generalization experiments. The matrix showcases the model's classification accuracy, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications, elucidating its ability to discern genuine videos from attacks.

patches were extracted from input videos and batched as inputs for our model. The deployment environment consisted of an Ubuntu Linux operating system, an Intel Xeon CPU, and a Nvidia Tesla T4 GPU augmented with the CUDA toolbox for PyTorch model training. Our findings reveal that, on average, our system requires approximately 40ms for processing a batch of 32 image patches, showcasing its commendable efficiency.

When presented with a video clip depicting a session of a phygital game, we suggest deploying our standalone spoof detection system once at the conclusion of a performance, concurrently with the execution of game logic. This approach serves to promptly identify any potential instances of spoofing attacks within the captured video footage.

1) COMPARATIVE ANALYSIS OF COMPUTATIONAL COMPLEXITY

To compare the computational complexity of our approach with other lightweight models, we conducted a detailed analysis of the Floating Point Operations Per Second (FLOPs) and model parameters. As depicted in Table 11, it's evident that our approach utilizing EfficientNet-B0 exhibits a higher number of model parameters and involves more operations, consequently resulting in lower inference time. In contrast, ShuffleNetv2 boasts the fewest parameters and requires fewer floating point operations, leading to faster inference speeds.

TABLE 11. Comparison of computational complexity metrics including Floating Point Operations Per Second (FLOPs) and model parameters for EfficientNet-B0, MobileNetV2, and ShuffleNetV2 architectures, showcasing their respective inference speeds and computational demands.

Method	GFLOPS	Inference speed (ms)	Params (M)
ShuffleNetv2 [39]	4.72	11.32	1.25
MobileNetV2 [40]	10.0	13.5	2.2
EfficientNet-B0 [51]	12.8	26	4.01

F. DISCUSSION

In this section, we delve into the comparative performance and computational characteristics of three lightweight architectures - EfficientNet-B0, MobileNetV2, and ShuffleNetV2, in the context of video-replay spoof detection using our Sqiller-Spoof dataset. Upon training and testing on our dataset, we observed that all models demonstrated relatively comparable performance in terms of both accuracy and HTER. However, our investigation into their generalization capabilities revealed a notable disparity. Our approach leveraging EfficientNet-B0 exhibited superior generalization abilities, while the method based on ShuffleNetV2 demonstrated the poorest performance in this regard.

Furthermore, our assessment of the computational complexity of these models provided valuable insights. ShuffleNetV2 showcased fewer floating point operations, resulting in higher inference speeds. Conversely, EfficientNet-B0

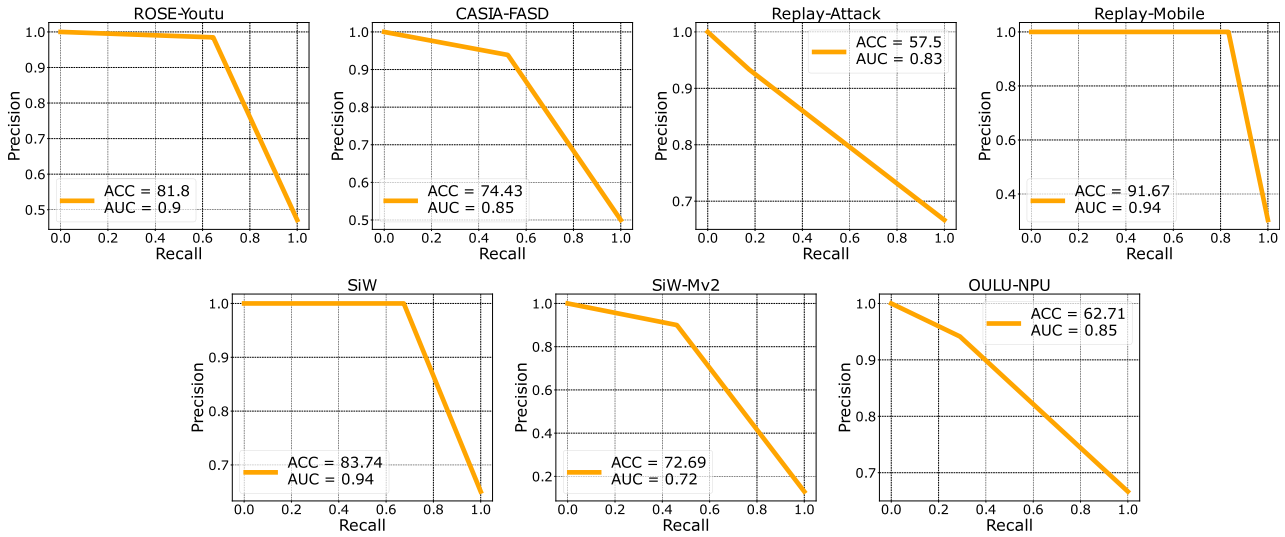

FIGURE 11. Precision-Recall curve depicting the performance of ShuffleNet V2 across different datasets employed for generalization experiments. The curve illustrates the trade-off between precision and recall at various classification thresholds, providing insights into the model's ability to accurately identify genuine videos while minimizing misclassifications.

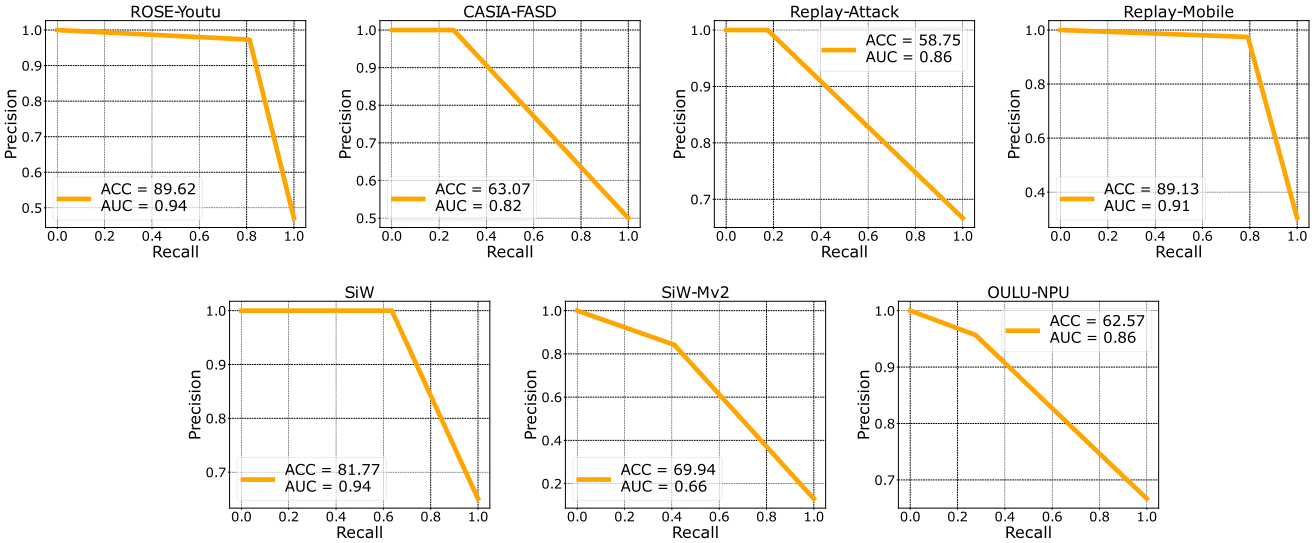

FIGURE 12. Precision-Recall curve showcasing the performance of MobileNetV2 across a range of datasets utilized for generalization experiments. The curve delineates the relationship between precision and recall at different classification thresholds, offering insights into the model's effectiveness in accurately detecting genuine videos amidst potential attacks.

incurred more floating point operations, leading to the least efficient inference speed among the three models.

These findings underscore the trade-offs inherent in selecting a suitable architecture. While lightweight architectures like ShuffleNetV2 offer computational efficiency and adaptability to specific domains, they may encounter challenges with overfitting and a lack of generalization to unseen scenarios. On the other hand, slightly larger architectures such as EfficientNet-B0 may introduce higher computational complexity but demonstrate superior performance in handling diverse and unseen scenarios.

Ultimately, the choice of model for a given application hinges on considerations related to hardware and software resources, as well as the specific requirements and constraints of the target deployment environment.

The proposed approach based on EfficientNet-B0 for video replay spoof detection in phygital games holds significant industrial relevance due to its combination of accuracy, efficiency, and adaptability. Firstly, EfficientNet-B0's ability to achieve high accuracy in detecting video replay spoof attacks enhances the security and integrity of phygital gaming environments. By accurately distinguishing between genuine player actions and fraudulent attempts, the approach ensures a fair and authentic gaming experience for users, mitigating the risk of cheating or manipulation. Secondly, the efficiency of EfficientNet-B0 makes it suitable for real-time deployment in

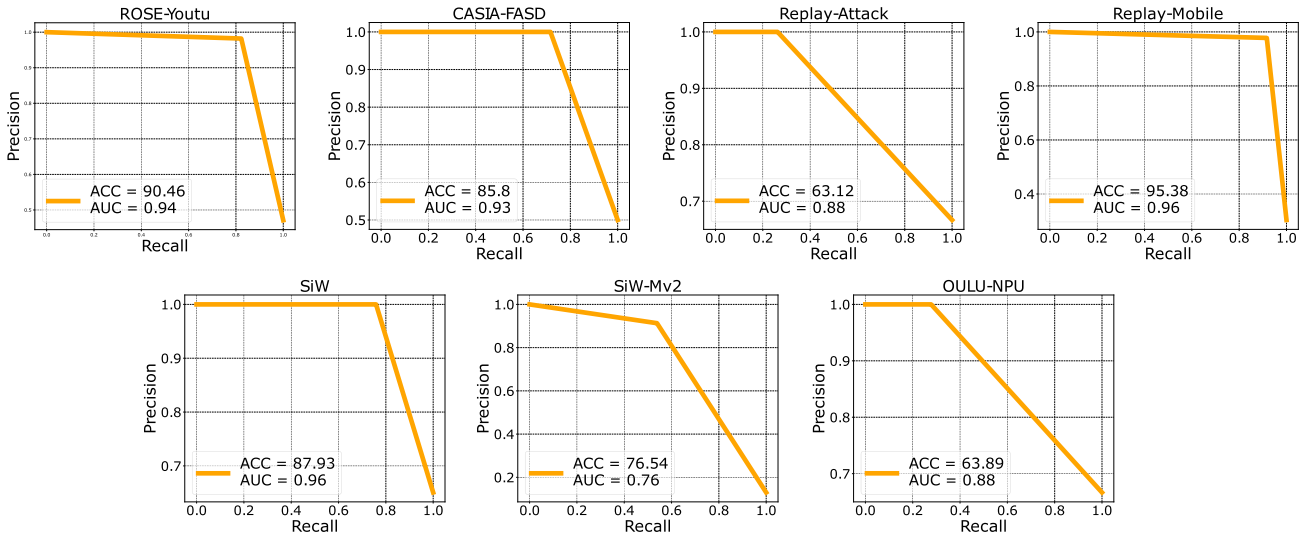

FIGURE 13. Precision-Recall curve presenting the performance of EfficientNet-B0 across diverse datasets employed for generalization experiments. The curve highlights the balance between precision and recall at varying classification thresholds, providing an indication of the model's ability to precisely identify genuine videos while minimizing false positives and false negatives.

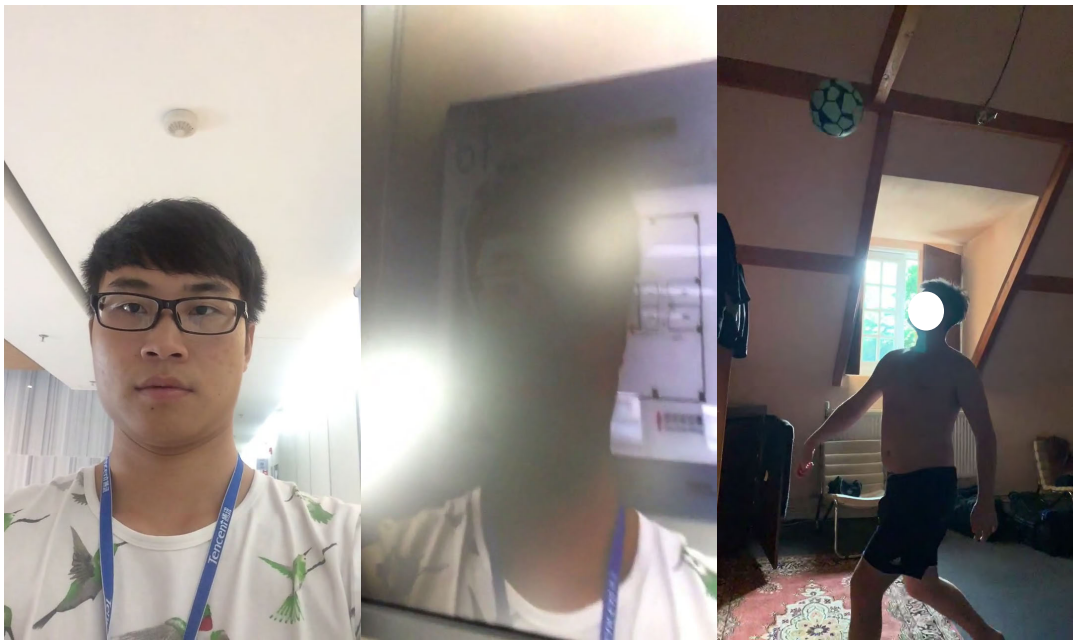

FIGURE 14. Illustrative instances where model limitations arise: Left - Glasses causing misclassification due to reflections. Middle - Reflections on device screens challenging classification. Right - Misinterpreting wooden beams as display bezels. These insights emphasize the need for context-specific model adjustments.

industrial applications. Its optimized architecture minimizes computational resources while maintaining robust performance, enabling seamless integration into existing phygital gaming systems without imposing significant overhead on hardware or infrastructure.

Moreover, the adaptability of EfficientNet-B0 allows for scalability and customization to accommodate diverse gaming scenarios and environments. Whether deployed in mobile applications or embedded within gaming consoles, the approach can be tailored to meet the specific requirements and constraints of different platforms and use cases.

Overall, the industrial significance of the proposed approach lies in its ability to enhance the security, fairness, and user experience of phygital games while offering efficient and adaptable solutions that align with the demands of modern gaming ecosystems. As the gaming industry continues to evolve and innovate, leveraging advanced techniques like EfficientNet-B0 for video replay spoof detection is

paramount to staying ahead of emerging threats and ensuring the integrity of gaming experiences.

G. LIMITATIONS

While our model demonstrated excellent performance in detecting video replay attacks during cross-validation studies, our subsequent exploration into generalization on face biometric recognition datasets revealed certain limitations. In cases where a real user wearing glasses is positioned close to the camera, our model occasionally misclassifies videos as attacks due to the presence of specular surfaces like glasses, which can reflect light and cause confusion (see Figure 14 (left)). However, it's important to note that these issues might have limited impact in the context of phygital games. This is because the game's rules dictate player positions and visibility requirements, ensuring our model's accuracy is maintained.

Similarly, our model might encounter challenges with videos where the device screen reflects external light, as depicted in Figure 14 (middle). Yet, this scenario is less relevant in the phygital games setting where gameplay logic itself might prevent these situations.

Another limitation arises when our model misinterprets wooden beams in the background as display bezels, classifying genuine videos as replay attacks, as shown in Figure 14 (right). Addressing this limitation might involve diverse training strategies. Furthermore, the accuracy values for the "attack with visible display bezels" class (Table 4) suggest potential overfitting concerns. To tackle this, exploring varied interpretations of display bezels during model training could be beneficial.

VI. CONCLUSION AND FUTURE WORK

Our effort to address video-replay spoofing detection in the context of phygital games has yielded valuable insights and practical solutions. By designing a specific dataset for this environment and employing a sophisticated model, we achieved impressive accuracy in spotting video replay attacks. Our study's focus on the distinct challenges of game settings, such as different camera angles, player interactions, and varying conditions, showcases a thorough security approach. Our innovative video-level prediction method, based on detailed frame analysis and Moiré pattern detection, has effectively distinguished genuine actions from fraudulent ones. Moreover, the model's resilience to different lighting and player positions makes it well-suited for real-world phygital gaming scenarios.

Nonetheless, we've identified certain limitations, especially with reflective surfaces, device screen reflections, and certain visual features. These insights guide us to consider specific enhancements for improved real-world performance. Importantly, our method performs well on datasets sharing the same video resolution as our training videos. However, adapting to different resolutions may require creating new datasets in the future. Our work advances security in the evolving domain of phygital games while providing valuable lessons about balancing technology and physical experience. As this field progresses, our contributions promise safer, more secure, and more engaging experiences, bridging the gap between the digital and physical worlds.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the International Federation of Teqball (FITEQ) for the support and for providing the necessary resources and infrastructure to collect the dataset used in our deep learning experiments. Their contributions have been invaluable in enabling the success of the research.

REFERENCES

[1] T. Murray, "The IOC remains focused on 'virtual sports' over esports," Esports Observer, Berlin, Germany, Germany. Accessed: Nov. 12, 2023. [Online]. Available: https://esportsobserver.com/ioc-virtual-sports-over-esports/

[2] *The Digital Football Game*, SQILLER App, FITEQ, Budapest, Hungary, 2019.

[3] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of Fourier spectra," in *Proc. SPIE*, vol. 5404, Orlando, FL, USA, 2004, pp. 296–303.

[4] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Proc. 11th Eur. Conf. Comput. Vis.*, Heraklion, Crete, Greece. Berlin, Germany: Springer, 2010, pp. 504–517.

[5] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[6] D. F. Smith, A. Wiliem, and B. C. Lovell, "Face recognition on consumer devices: Reflections on replay attacks," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 736–745, Apr. 2015.

[7] K. Zuiderveld, *Contrast Limited Adaptive Histogram Equalization*. New York, NY, USA: Academic, 1994.

[8] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 3557–3560.

[9] J. Bai, T.-T. Ng, X. Gao, and Y.-Q. Shi, "Is physics-based liveness detection truly possible with a single image?" in *Proc. IEEE Int. Symp. Circuits Syst.*, Paris, France, 2010, pp. 3425–3428.

[10] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Washington, DC, USA, Oct. 2011, pp. 1–7.

[11] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using texture and local shape analysis," *IET Biometrics*, vol. 1, no. 1, p. 3, 2012.

[12] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *Proc. Int. Conf. Biometrics (ICB)*, Madrid, Spain, Jun. 2013, pp. 1–6.

[13] N. Kose and J.-L. Dugelay, "Countermeasure for the protection of face recognition systems against mask attacks," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, Apr. 2013, pp. 1–6.

[14] N. Kose and J.-L. Dugelay, "Shape and texture based countermeasure to protect face recognition systems against mask attacks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Shanghai, China, Jun. 2013, pp. 111–116.

[15] J. Galbally and S. Marcel, "Face anti-spoofing based on general image quality assessment," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 1173–1178.

[16] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 710–724, Feb. 2014.

[17] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Montreal, QC, Canada, Sep. 2015, pp. 2636–2640.

[18] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1818–1830, Aug. 2016.

[19] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014, *arXiv:1408.5601*.
[20] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Proc. Chin. Conf. Biometric Recognit.*, Chengdu, China. Cham, Switzerland: Springer, 2016, pp. 611–619.
[21] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Oulu, Finland, Dec. 2016, pp. 1–6.
[22] C. Nagpal and S. R. Dubey, "A performance evaluation of convolutional neural networks for face anti spoofing," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[23] H. Xue, J. Ma, and X. Guo, "A hierarchical multi-modal cross-attention model for face anti-spoofing," *J. Vis. Commun. Image Represent.*, vol. 97, Dec. 2023, Art. no. 103969.
[24] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13868–13882, 2017.
[25] DMJ Tax, "One-class classification; concept-learning in the absence of counter-examples," Ph.D. thesis, ASCI Dissertation Series 65, Delft Univ. Technol., Delft, Delft, The Netherlands, 2001.
[26] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing," in *Proc. Int. Conf. Biometrics (ICB)*, Gold Coast, QLD, Australia, Feb. 2018, pp. 75–81.
[27] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4675–4684.
[28] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 389–398.
[29] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 290–306.
[30] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *Proc. Int. Conf. Biometrics (ICB)*, Crete, Greece, Jun. 2019, pp. 1–8.
[31] X. Li, W. Wu, T. Li, Y. Su, and L. Yang, "Face liveness detection based on parallel CNN," *J. Phys., Conf. Ser.*, vol. 1549, no. 4, Jun. 2020, Art. no. 042069.
[32] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5294–5304.
[33] A. Benlamoudi, S. E. Bekhouche, M. Korichi, K. Bensid, A. Ouahabi, A. Hadid, and A. Taleb-Ahmed, "Face presentation attack detection using deep background subtraction," *Sensors*, vol. 22, no. 10, p. 3760, May 2022.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[35] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2704–2713.
[36] A. Gordon, E. Eban, O. Nachum, B. Chen, H. Wu, T.-J. Yang, and E. Choi, "MorphNet: Fast & simple resource-constrained structure learning of deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1586–1595.
[37] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[38] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[39] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.
[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[41] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.*, Urumqi, China. Cham, Switzerland: Springer, 2018, pp. 428–438.
[42] Y. Martindez-Díaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza, "ShuffleFaceNet: A lightweight face architecture for efficient and highly-accurate face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2721–2728.
[43] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su, "VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2647–2654.
[44] Y. Martínez-Díaz, M. Nicolás-Díaz, H. Méndez-Vázquez, L. S. Luevano, L. Chang, M. Gonzalez-Mendoza, and L. E. Sucar, "Benchmarking lightweight face architectures on specific face recognition scenarios," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 6201–6244, Dec. 2021.
[45] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano, M. Nicolás-Díaz, L. Chang, and M. González-Mendoza, "Towards accurate and lightweight masked face recognition: An experimental evaluation," *IEEE Access*, vol. 10, pp. 7341–7353, 2022.
[46] A. Rosenfeld, *Digital Picture Processing*. Cambridge, MA, USA: Academic, 1976.
[47] X. Cheng, Z. Fu, and J. Yang, "Multi-scale dynamic feature encoding network for image demoiréing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, 2019, pp. 3486–3493.
[48] C. Yang, Z. Yang, Y. Ke, T. Chen, M. Grzegorzek, and J. See, "Doing more with Moiré pattern detection in digital photos," *IEEE Trans. Image Process.*, vol. 32, pp. 694–708, 2023.
[49] B. He, C. Wang, B. Shi, and L.-Y. Duan, "FHDE^{2}Net: Full high definition demoireing network," in *Proc. Comput. Vis. ECCV 16th Eur. Conf.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 713–729.
[50] V. D. Huszár and V. K. Adhikarla, "Live spoofing detection for automatic human activity recognition applications," *Sensors*, vol. 21, no. 21, p. 7339, Nov. 2021.
[51] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 6105–6114.
[52] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2012, pp. 1–7.
[53] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, New Delhi, India, Mar. 2012, pp. 26–31.
[54] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.
[55] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2268–2283, Oct. 2016.
[56] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2016, pp. 1–7.
[57] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 612–618.
[58] X. Tu, H. Zhang, M. Xie, Y. Luo, Y. Zhang, and Z. Ma, "Deep transfer across domains for face antispoofing," *J. Electron. Imag.*, vol. 28, no. 4, 2019, Art. no. 043001.
[59] 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 Trans. Inf. Forensics Security*, vol. 15, pp. 42–55, 2020.
[60] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1178–1186.

[61] K. Kotwal, S. Bhattacharjee, P. Abbet, Z. Mostaani, H. Wei, X. Wenkang, Z. Yaxi, and S. Marcel, "Domain-specific adaptation of CNN for detecting face presentation attacks in NIR," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 1, pp. 135–147, Jan. 2022.
[62] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *Proc. IEEE 6th Int. Conf. Biometrics: Theory, Appl. Syst. (BTAS)*, Washington, DC, USA, Sep. 2013, pp. 1–8.
[63] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
[64] L. Li, Z. Gao, L. Huang, H. Zhang, and M. Lin, "A dual-modal face anti-spoofing method via light-weight networks," in *Proc. IEEE 13th Int. Conf. Anti-Counterfeiting, Secur., Identificat. (ASID)*, Xiamen, China, Oct. 2019, pp. 70–74.
[65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.
[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Intl. Conf. Learning Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.
[67] *ISO/IEC JTC 1/SC 37 Biometrics Information Technology—Biometric Presentation Attack Detection Part 3: Testing and Reporting*, ISO, International Publisher for Standardization, Geneva, Switzerland, Sep. 2017.
[68] *Accuracy (Trueness and Precision) of Measurement Methods and Results*, Organització Internacional per a la Normalització, Int. publisher Standardization, Geneva, Switzerland, 1994.
[69] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[70] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano, and M. Gonzalez-Mendoza, "Exploring the effectiveness of lightweight architectures for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 6391–6401.
[71] Z. Çamlica, H. R. Tizhoosh, and F. Khalvati, "Medical image classification via SVM using LBP features from saliency-based folded data," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 128–132.
[72] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2005, pp. 886–893.
[73] Y. Zhang, R. K. Dubey, G. Hua, and V. L. L. Thing, "Face spoofing video detection using spatio-temporal statistical binary pattern," in *Proc. TENCON IEEE Region 10 Conf.*, Jeju Island, South Korea, Oct. 2018, pp. 0309–0314.
[74] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Can face anti-spoofing countermeasures work in a real world scenario?" in *Proc. Int. Conf. Biometrics (ICB)*, Madrid, Spain, 2013, pp. 1–8.
[75] T. P. Nguyen, N.-S. Vu, and A. Manzanera, "Statistical binary patterns for rotational invariant texture classification," *Neurocomputing*, vol. 173, pp. 1565–1577, Jan. 2016.
[76] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "LBP–top based countermeasure against face spoofing attacks," in *Proc. Comput. Vis. ACCV Workshops*, J.-I. Park and J. Kim, Eds. Berlin, Germany: Springer, 2013, pp. 121–132.
[77] *Deepstream Sdk*, NVIDIA, Santa Clara, CA, USA, 2023.
[78] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

VIKTOR DÉNES HUSZÁR (Member, IEEE) is currently a Researcher with the Doctoral School of Military Engineering, National University of Public Service. His research interests include computer vision and artificial intelligence. He is a member of Hungarian Economic Association. He received several Hungarian and international awards, including the Industrial Innovation Award, the Red Dot Award, and the IF Design Award. He serves as the Chairperson for FITEQ, the governing body of Teqball, and the Head of the Digital Committee of FINA, the governing body of aquatics sports. He is also an international speaker on computer vision-based technologies.

VAMSI KIRAN ADHIKARLA received the dual master's (M.Tech. and M.Sc.) degree from Blekinge Tekniska Högskola, Sweden, and JNTU Hyderabad, India, in 2011, and the Ph.D. degree in information science from the Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Hungary, in 2015. From 2012 to 2015, he was a Marie Curie Early Stage Researcher with Holografika, Hungary. From 2015 to 2017, he was a Postdoctoral Researcher with the Department of Computer Graphics, Max-Planck-Institute for Informatics. He is currently a Researcher with the International Federation of Teqball, Hungary. His research interests include real-time computer vision, deep learning for visual analysis, 3D computer vision, and 3DTV. In 2021, he was awarded the Marie Skłodowska-Curie Individual Fellowship Award.