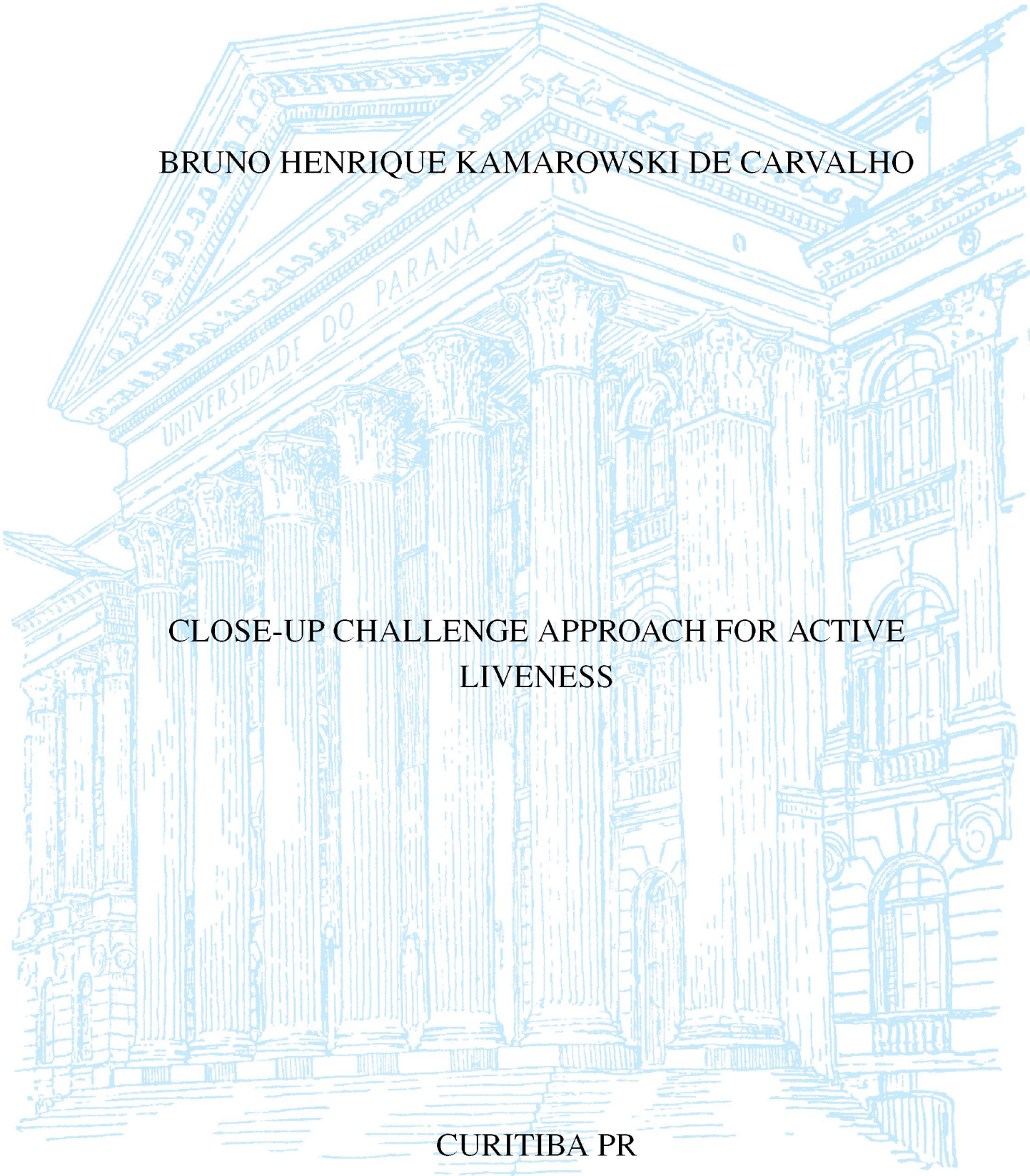
UNIVERSIDADE FEDERAL DO PARANÁ

BRUNO HENRIQUE KAMAROWSKI DE CARVALHO



CLOSE-UP CHALLENGE APPROACH FOR ACTIVE LIVENESS

CURITIBA PR
2024

BRUNO HENRIQUE KAMAROWSKI DE CARVALHO

CLOSE-UP CHALLENGE APPROACH FOR ACTIVE LIVENESS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática, no Programa de Pós-Graduação em Informática, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: David Menotti Gomes.

Coorientador: Roger Leitzke Granada.

CURITIBA PR
2024

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Carvalho, Bruno Henrique Kamarowski de
Close-up challenge approach for active liveness / Bruno Henrique Kamarowski de Carvalho. – Curitiba, 2024.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática.

Orientador: David Menotti Gomes
Coorientador: Roger Leitzke Granada

1. Reconhecimento facial (Computação). 2. Conjunto de caracteres (Processamento de dados). 3. Falsificação. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Informática. III. Gomes, David Menotti. IV. Granada, Roger Leitzke. V. Título.

Bibliotecário: Leticia Priscila Azevedo de Sousa CRB-9/2029

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **BRUNO HENRIQUE KAMAROWSKI DE CARVALHO**, intitulada: **Close-up challenge approach for active liveness**, sob orientação do Prof. Dr. DAVID MENOTTI GOMES, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 26 de Março de 2025.

Assinatura Eletrônica
26/03/2025 14:09:39.0
DAVID MENOTTI GOMES
Presidente da Banca Examinadora

Assinatura Eletrônica
26/03/2025 11:06:33.0
RAYSON BARTOSKI LAROCA DOS SANTOS
Avaliador Externo (PONTIFICIA UNIVERSIDADE CATOLICA DO PARANÁ- PUCPR)

Assinatura Eletrônica
28/03/2025 17:29:39.0
ANDRÉ RICARDO ABED GRÉGIO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
26/03/2025 11:01:55.0
ROGER LEITZKE GRANADA
Coorientador(a) (UNICO IDTECH)

Dedico esse trabalho à minha querida irmã, que tanto admiro e que sempre esteve ao meu lado.

ACKNOWLEDGEMENTS

Ao professor David Menotti e Roger Granada, por terem me guiar durante toda a realização deste trabalho. Aos meus familiares por me incentivarem a me dedicar aos estudos e me concederem condições para tal. Aos meus amigos pelas palavras de apoio e por todos os momentos juntos.

Também agradeço à UFPR, à FUNPAR e à Unico - IDTech por financiarem e viabilizarem o projeto associado ao desenvolvimento deste trabalho.

RESUMO

A tarefa de Anti-Spoofing Facial (FAS) foca na detecção de tentativas de enganar sistemas de autenticação facial. A grande maioria dos estudos nessa área se concentra em abordagens passivas, que não exigem nenhuma interação especial por parte do usuário. Em contraste, o subcampo da detecção ativa de vivacidade — onde a autenticidade é verificada por meio de ações realizadas pelo usuário — permanece pouco explorado. Essa lacuna se deve principalmente à falta de conjuntos de dados públicos adequados para tarefas de vivacidade ativa, o que leva a pesquisas irreproduzíveis, métodos desatualizados e análises comparativas limitadas. Para enfrentar esses problemas, este trabalho apresenta um novo conjunto de dados de vivacidade ativa com ênfase em vídeos com movimentos em close-up. O conjunto contém 714 amostras genuínas coletadas de voluntários e 1.847 amostras falsas criadas usando imagens do CelebA e vídeos do CelebV exibidos em diferentes instrumentos de ataque de apresentação. Propomos quatro protocolos de avaliação para testar a capacidade de generalização de modelos de detecção de vivacidade ativa em cenários desafiadores, como lidar com ataques desconhecidos, instrumentos não vistos anteriormente e variações nos padrões de aquisição das câmeras. Além disso, este trabalho apresenta um novo modelo baseado em trabalhos anteriores, que integra invariantes projetivos com embeddings faciais para uma extração de características mais robusta. Essa abordagem melhora significativamente as técnicas existentes, superando outras linhas de base na detecção de tentativas de falsificação.

Palavras-chave: Vivacidade Facial Ativa, Dataset de Close-Up, Face anti-spoofing.

ABSTRACT

The Face Anti-Spoofing (FAS) task focuses on detecting attempts to deceive facial authentication systems. The vast majority of studies in this field focus on passive approaches, which do not require any special interaction with the user. In contrast, the subfield of active liveness detection — where authenticity is verified through user-performed actions — remains underexplored. This gap is primarily due to a lack of public datasets suitable for active liveness tasks, leading to irreproducible research, outdated methods, and limited comparative analysis. To address these issues, this work introduces a new active liveness dataset emphasizing videos with close-up movements. The dataset consists of 714 genuine samples collected from volunteer subjects and 1,847 spoof samples created using CelebA images and CelebV videos displayed across various presentation attack instruments. We propose four evaluation protocols to assess the generalization capabilities of active liveness detection models in challenging scenarios, such as handling unknown attacks, unseen instruments, and variations in camera acquisition patterns. Additionally, this work presents a new model that builds on prior work, integrating projective invariants with facial embedding for more robust feature extraction. This approach significantly improves upon existing techniques, outperforming other baselines in detecting spoofing attempts.

Keywords: Active Face Liveness, Close-Up Dataset, Face anti-spoofing.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

PA	Presentation attack
PAI	Presentation attack instrument
PAD	Presentation attack detector
FAS	Face anti-spoofing
APCER	Attack presentation classification error rate
BPCER	Bona fide presentation classification error rate
ACER	Average classification error rate
FAR	False acceptance rate
FRR	False rejection rate
CNN	Convolutional neural network
RNN	Recurrent neural network
LBP	Local binary pattern
rPPG	remote photoplethysmography
FPS	Farthest Point Sampling
GPU	Graphics processing unit

SUMMARY

Chapter 1

INTRODUCTION

Over the decades, the evolution of human-machine interfaces has been notable, with systems evolving from mechanical schemes to touch interfaces, voice commands, and even gesture-based controls. Following this trend, authentication mechanisms began to explore biometric information, popularizing facial authentication methods among various forms of biometrics(Jain et al. (2006)). Currently, these methods are employed in various infrastructures, such as entryways, digital banking, and device screen locks(Rui e Yan (2018)).

Parallel to the evolution of facial authentication, problems associated with exploiting vulnerabilities in these mechanisms have emerged, typically occurring in two ways: through injection attacks or presentation attacks. The former happens when the attacker bypasses the app's security measures and inserts their own media instead of the one from the device's camera. The latter, as the name suggests, occurs when an object or a mischaracterized face is shown to the device's camera, typically to assume the identity of someone else or to obfuscate the attacker's identity (Yu et al. (2023)). Several instruments are used to impersonate victims, such as high-quality printed photos of targets, social media videos displayed on digital screens, handcrafted or 3D-printed masks, and others (Kumar et al. (2017)). The instrument used to perform impersonation attacks is called a presentation attack instrument (PAI) (Ming et al. (2020)).

To identify these types of attacks, it is necessary to implement methods that perform the task of facial liveness verification, also known as presentation attack detection (PAD) when limited to presentation attacks (Marcel et al. (2019)). Methods that perform this task can also be referred to as face anti-spoofing (FAS) and mostly consist of binary classifiers labeling inputs as spoof when the media contains a liveness attack or live otherwise (Raheem et al. (2019)).

FAS can be broadly categorized into two types: active and passive methods. *Passive methods* detect signs of vitality from facial images or videos without requiring any explicit user interaction. In contrast, *active methods* require users to perform specific actions or gestures during authentication to confirm their presence (Yu et al. (2023)). We will focus on the latter approach since it is in the main scope of the research.

Within the realm of active liveness, a wide array of user interactions can be used. To name a few, it is possible to mention methods based on the measurement of physiological signals, spontaneous facial movements, information injection into the sample, and challenge-response methods (Antil e Dhiman (2025)). It is noteworthy that the common characteristic of all these approaches is their dynamic aspect; that is, the information exploited is contained over time as the system interacts with the user. Therefore, FAS mechanisms that adopt this approach are also known as dynamic methods, as opposed to static methods, which rely on the information contained in a single image to perform liveness verification (Yu et al. (2020a)).

1.1 MOTIVATION

The increasing prevalence and sophistication of face spoofing attacks targeting face recognition systems is one of the main reasons for the continuous research and proposals regarding face liveness. As these systems become integral to various applications, such as financial transactions, access control, and personal device security, ensuring their robustness against such attacks is paramount.

It is believed that active liveness detection techniques offer a more resilient defense by analyzing dynamic responses to specific prompts. This dynamic interaction provides richer data for distinguishing between genuine users and spoofing attempts, making it harder for attackers to deceive the system using static images or pre-recorded videos.

In light of this situation, this work aims to explore active face liveness approaches based on the close-up challenge-response paradigm. By mapping meaningful features and enhancing the performance of existing proposals.

1.2 CHALLENGES

In the literature on passive facial liveness, a wide range of methods and implementations can be found, as well as the availability of different public datasets that can be used for experimentation and performance comparison of various methods.

However, the scenario for active liveness is significantly different. Primarily, the scarcity of public data for conducting research in this area is by far the most alarming challenge, as up to the present moment, there is no public dataset designed to study active face liveness. This hinders the execution of studies in the field and reveals the rudimentary stage of research addressing active liveness. Moreover, works on the topic often create their own datasets and report results on these private data, making it impossible to fairly compare different approaches.

Additionally, the few studies that explore the challenge-response approach based on close-ups do not directly incorporate raw image information into their models. Instead, they typically rely on extracting handcrafted features from a few frames of the input media. The lack of methods utilizing learned maps to extract information directly from images presents a challenge, as it represents an unexplored and innovative approach.

1.3 RESEARCH HYPOTHESIS

We hypothesize that the performance of existing Close-Up-based active FAS methods can be enhanced by integrating movement-based features with texture and spatial information.

1.4 PROPOSED APPROACH

The proposed approach revolves around two main points: the creation of an active dataset for facial liveness and the exploration of a novel method by adding spatial embedding information to other well-known methods.

The active dataset utilizes a close-up interaction that involves asking the user to align their face within a region shown on the capture device display. Initially, the user is positioned far from the device's camera. After a few seconds, the alignment region changes, requiring the user to move closer to fit their face into the area of interest. Figure 1.1 illustrates this interaction.

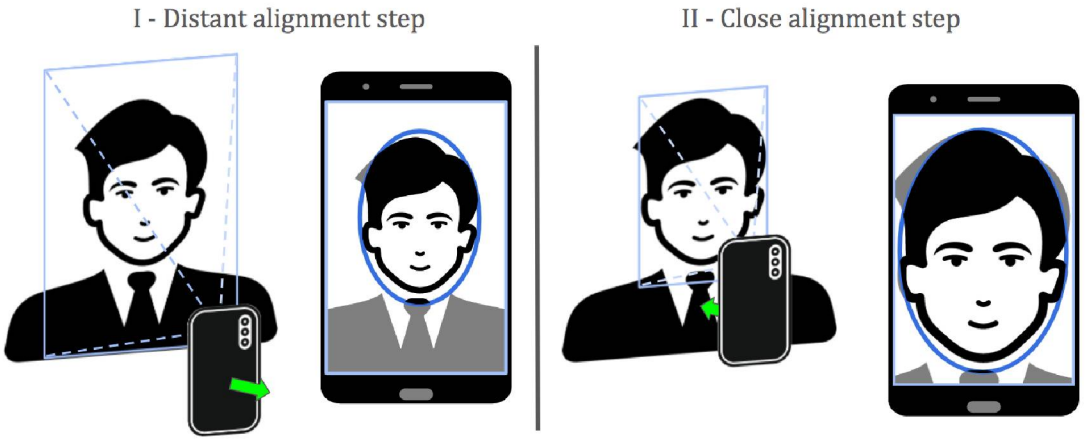


Figure 1.1: Close-Up movement scheme. First, faces are captured far from the camera by positioning them within a small region displayed on the screen, and then they are captured close, by fitting within a larger region shown.

The few active methods that employ the close-up challenge-response approach typically use only high-level features based on the concept of projection invariants, as defined by Riccio e Dugelay (2007). In this work, we propose a method following a hybrid architecture, combining these commonly used features with embedding information extracted from images by well-known encoders in the literature.

1.5 OBJECTIVES

This work aims to create a diverse and challenging dataset for training active methods based on the close-up interaction, as well as propose evaluation protocols to serve as a benchmark for related works in the field.

Furthermore, it proposes a new active presentation attack detector by adding spatial embedding information on existing models, comparing the results with the latest active benchmarks based on the close-up challenge.

1.6 PUBLICATIONS

Sub-products of this proposal were published in previous works:

- Conference on Graphics, Patterns and Images (SIBGRAPI) - Workshop of Works in Progress (WIP) - Multi-challenge database for active liveness - Kamarowski et al. (2023)
- Conference on Graphics, Patterns and Images (SIBGRAPI) - Workshop of Works in Progress (WIP) - Hybrid method for active face anti-spoofing based on close-up challenge Kamarowski et al. (2024)
- Scientific Reports - Hybrid Close-up approach for new Active Face liveness benchmark - To be Submitted

1.7 CONTRIBUTIONS

The contributions of this work are twofold:

- Establishment of a common dataset, enabling fair and standardized comparison of works in the field.
- Creation of a new active method based on the close-up challenge-response interaction combining well-known motion-based distortion features with spatial information.

1.8 OUTLINE

The remainder chapters in this work are organized as follows. Chapter 2 describes fundamental concepts and metrics related to liveness detection. Chapter 3 presents a study of related work. Chapter 4 presents this work's proposed approach, which is evaluated as described and presented in Chapter 5. Limitations and possibilities for future work are finally discussed in Chapter 6.

Chapter 2

THEORETICAL BACKGROUND

Relevant concepts and metrics related to the field of face liveness detection and networks used in Chapters 4 and 5 are now presented. A foundational understanding of machine learning is expected from the reader. This chapter starts by presenting the concept of Face Spoofing Attacks along with a general description of evaluation metrics and a brief overview of each method used as a baseline in this study.

2.1 FACE SPOOFING ATTACKS

A facial spoofing attack is an attempt to interfere with the operation of a facial biometric system. This type of attack can occur in two main ways: either during or after the capture of the media used by the biometric system. Attacks occurring after capture are known as injection attacks. In these cases, the attacker bypasses the media acquisition system by inserting manipulated data that could potentially be used by the biometric system. On the other hand, attacks that take place during the capture process are referred to as presentation attacks (PA). Here, the attacker exposes a presentation attack instrument (PAI) to the camera at the time of media capture.

Presentation attacks can be further divided into two categories: obfuscation and impersonation attacks. In an obfuscation attack, the attacker's goal is simply to avoid recognition by the biometric system. This type of attack typically uses PAIs to intentionally obscure or make facial features unrecognizable. Impersonation attacks, however, target a specific victim. PAIs in impersonation attacks typically employ representations of the victim's face. Only impersonation attacks fall within the scope of this work, and thus, the term "presentation attack" (PA) will hereafter refer specifically to impersonation attacks.

Various PAIs can be used to assume the identity of a victim, and it is common in the literature to categorize presentation attacks by the instrument employed. The most common attack types in the literature include photo attacks, which use a printed image; display attacks, which use a static image on a digital screen; replay attacks, which involve video, GIFs, or any media with dynamic characteristics; and mask attacks, which use masks to carry out the spoof. In all cases, the instrument can vary greatly in complexity: it may be very simple, such as printed photos on a standard printer, or highly complex, such as synthetic masks that accurately represent the target's facial features.

2.2 EVALUATION METRICS

The task of liveness verification is typically modeled as a binary classification problem with two classes: spoof and bonafide (often referred to as live). Generally, datasets are heavily imbalanced toward the spoof class to ensure broad representation across various presentation attacks (PA) and presentation attack instruments (PAIs). Consequently, metrics that individually report the predictions for each class are commonly used.

Following the evaluation protocol established by ISO/IEC 30107-3 (ISO/IEC, 2023), presentation attack detectors should report the bonafide presentation classification error rate (BPCER) and the attack presentation classification error rate (APCER) for each PA. In the literature, it is also common to report the average classification error rate (ACER) in intra-dataset experiments.

The APCER for a given PA and the BPCER are defined as follows

$$APCER_{PA} = 1 - \frac{1}{N_{PA}} \sum_{i=1}^{N_{PA}} Res_i, \tag{2.1}$$

$$BPCER = \frac{1}{N_{BF}} \sum_{i=1}^{N_{BF}} Res_i, \tag{2.2}$$

where N_{PA} and N_{BF} represent the number of spoof samples for that PA and the number of bona fide samples, respectively. Res_i takes the value 1 if the i-th presentation is classified as an attack presentation and 0 if classified as a bona fide presentation.

Finally, ACER, which is not mentioned in the ISO/IEC 30107-3 but is widely adopted in scientific studies on liveness, is defined by Boulkenafet et al. (2017b) as the average of the highest APCER value among all PAs and the BPCER, as follows

$$ACER = \frac{\max_{p \in PA}(APCER_p) + BPCER}{2}. \tag{2.3}$$

Except for the particularity of computing APCER for each PA, it is noteworthy that the BPCER and APCER metrics are analogous to the false rejection rate (FRR) and false acceptance rate (FAR), commonly used in other machine learning contexts.

2.3 BASELINES

The methods Camera Close-Up (Castelblanco et al., 2022) and Face Close-Up (Li et al., 2019), used as baselines for this work, are now presented. Both methods are based on the principle of invariant projections defined by Riccio e Dugelay (2007), calculated through distances between landmarks. To the best of our knowledge, these are the most recent active methods utilizing the close-up challenge.

2.3.1 Face Close-up

Face Close-Up is a method for facial liveness detection that includes three main modules: the Video Frame Selector, the Distortion Feature Extractor, and the Liveness Classifier. The first module begins by processing input video and selecting frames based on the size of the face in each frame, ensuring that the video sequence captures the face at various distances as the camera

moves closer or farther from it. Using a face detection algorithm, this module determines face sizes and selects frames that best represent these changes in distance.

With these frames, the Distortion Feature Extractor module detects 66 facial landmarks, including the chin, eyes, and lips, and calculates pairwise distances between these landmarks to capture facial geometry. Figure 2.1 shows a representation of the used landmarks. For each frame, these distances are normalized relative to a reference frame, which is established during a user registration phase. The result is a set of relative distances that reflect how facial geometry changes over time, forming a matrix that captures distortions across frames. This matrix serves as the core feature set for detecting liveness.

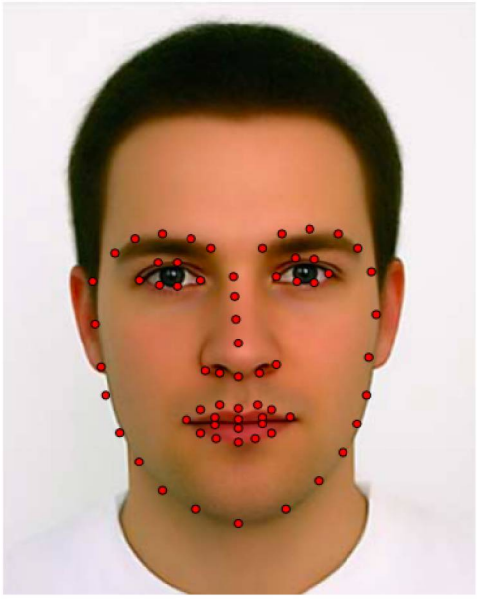


Figure 2.1: Landmarks used to extract distortions over time. Source: Li et al. (2019)

Finally, the Livenes Classifier module uses a convolutional neural network (CNN) to analyze this matrix of distortion features. The CNN architecture includes two convolutional layers and two pooling layers, followed by two fully connected layers, which help extract high-level patterns in the facial features to distinguish real faces from spoofed ones. The final output layer estimates the likelihood of the face being genuine or spoofed by providing probability estimates for each class.

2.3.2 Camera Close-up

Unlike the previous method designed for controlled environments, this approach is adapted for in the wild conditions where there is no fixed camera-user distance. Camera Close-Up is designed for both liveness detection and face verification, leveraging changes in facial geometry caused by natural camera movement and perspective shifts but only the FAS is explored in our work.

The preprocessing pipeline extracts a set of frames with detectable facial landmarks. Each frame's signature is calculated as a vector of Euclidean distances between all landmark pairs. These frame signatures are concatenated to form a video signature matrix, which is then normalized relative to a reference frame located at the video's midpoint, creating a matrix representing relative landmark distances.

To ensure consistent input dimensions regardless of video length, the method samples a fixed number of normalized frame signatures using stratified sampling. This approach selects frames equitably across the video's timeline to capture perspectives from close and far distances.

For classification, the method uses a neural network (NN) inspired by prior work but optimized for this paradigm's unique features. The NN model includes an initial custom convolutional kernel to highlight horizontal changes in landmark distances, followed by three convolutional layers with batch normalization and max pooling. Finally, the model employs three fully connected layers with ReLU activation, ending with a sigmoid neuron to classify videos as live or spoofed. This model architecture leverages landmark distance changes to robustly detect liveness and verify identity in natural, uncontrolled settings. Figure 2.2 summarizes the described architecture.

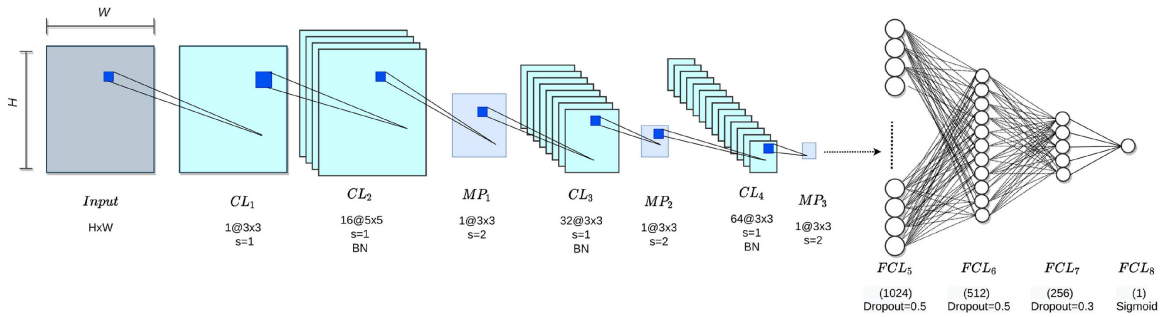


Figure 2.2: Camera Close-Up CNN architecture CL: 2D Convolutional layer; MP: 2D Max-Pooling; FCL: Fully connected layer; BN: Batch Normalization. Source: Castelblanco et al. (2022).

Chapter 3

RELATED WORK

In this section, we present published works related to FAS. Section 3.1 discusses important passive liveness datasets, Section 3.2 summarizes the passive face anti-spoofing methods studied, and Subsection 3.3 presents the active face anti-spoofing methods explored.

3.1 DATASETS

Advancements in passive liveness detection have led to the creation of various datasets tailored for this task. Recent studies consider not only the number of individuals and types of attacks but also the variety of scenarios, lighting conditions, camera quality, and the diversity of individuals included. Table 3.1 summarizes some of the most popular datasets in the literature, including data from the Face Close-up (Li et al., 2019) and Camera Close-up (Castelblanco et al., 2022) methods.

However, it is important to note that the cited passive datasets are not suitable for active approaches using the close-up paradigm, as they lack any form of interaction. An exception is the SiW dataset (Liu et al., 2018a), which includes a partition of images featuring individuals performing non-trivial tasks to assess the robustness of passive methods to variations in pose and distance. It is important to emphasize that this partition was created to stress-test the liveness verification capabilities of passive models in atypical scenarios. Due to this, the protocols suggested in this work do not meet the requirements of active methods and also have a limited number of samples performing the interaction of interest.

Moreover, the studies mentioned in Section 3.3 did not disclose the data used in their respective experiments. Therefore, to the best of our knowledge, there are no publicly available datasets designed specifically for active liveness detection using the close-up paradigm.

3.2 PASSIVE METHODS

Over the years, methods for Face Anti-Spoofing (FAS) have significantly advanced, shifting from simple handcrafted feature detection to the learning of feature maps through sophisticated techniques. Initially, researchers like Boulkenafet et al. (2017a) described facial appearance by applying Fisher vector encoding to features extracted from various color spaces. Similarly, Chingovska et al. (2012) assessed the effectiveness of texture features based on Local Binary Patterns (LBP) and their variations for classification.

In more recent developments, hybrid approaches have emerged, which combine handcrafted features with those extracted using deep neural networks such as in de Freitas Pereira

Table 3.1: Studied datasets' main characteristics.

Dataset	Citation	Samples	Subjects	Attack types	User interaction
NUAA	Tan et al. (2010)	5105 real, 7509 spoof	15	1	Passive
PRINT-ATTACK	Anjos e Marcel (2011)	200 real, 200 spoof	50	1	Passive
CASIA	Zhang et al. (2012)	150 real, 450 spoof	50	3	Passive
Replay-Attack	Chingovska et al. (2012)	200 real, 1000 spoof	50	3	Passive
MSU-MFSD	Wen et al. (2015)	110 real, 330 spoof	55	3	Passive
MSU-USSA	Patel et al. (2016)	1140 real, 9120 spoof	1140	2	Passive
MLFP	Agarwal et al. (2017)	150 real, 1200 spoof	10	2	Passive
Oulu-NPU	Boulkenafet et al. (2017b)	990 real, 3960 spoof	55	4	Passive
SiW	Liu et al. (2018a)	1320 real, 3300 spoof	165	6	Multiple angles, face expressions and subject movement
SiW-M	Liu et al. (2019a)	660 real, 968 spoof	493	13	Passive
HQ-WMCA	Mostaani et al. (2020)	555 real, 2349 spoof	51	10	Passive
DMAD	Wang et al. (2020)	900 real, 1800 spoof	300	6	Passive
Celeb A-Spoof	Zhang et al. (2020b)	156 384 real, 469 153 spoof	10 177	6	Passive
WFAS	Wang et al. (2023)	529 571 real, 853 729 spoof	469 920	18	Passive
Face Close-up dataset	Li et al. (2019)	710 real, 4970 spoof	71	3	Close up
Camera Close-up dataset	Castelblanco et al. (2022)	89 real, 2537 spoof	41	5	Close up
UFPR Close-Up	This work	391 real, 1043 spoof	714 live, 1847 spoof	5	Close up

et al. (2013) and Komulainen et al. (2013). Additionally, methods described in Liu et al. (2021), Garg et al. (2020), and Yu et al. (2020b) use traditional deep learning techniques, employing end-to-end Convolutional Neural Networks (CNNs) to map face images directly to liveness labels with direct supervision for training. In contrast, Zheng et al. (2021), Wang et al. (2021b), and Zhang et al. (2020a) adopt pixel-wise supervision.

Advanced deep learning methods go even further by striving for robustness against variations in input sensors and attack types. Following this trend, the methods Wang et al. (2022), Sun et al. (2023) and Le e Woo (2024) focus on domain generalization, where the model is trained only once. Another approach is domain adaptation, which involves adjusting the model using test data (Li et al., 2018; Wang et al., 2021a). Moreover, George e Marcel (2021) and Quan et al. (2021) aim to generalize to unseen attack types through zero-shot and few-shot learning strategies, which use little to no training data. Lastly, there is also anomaly detection, where the model learns accurate representations of live samples instead of spoof characteristics as demonstrated in Liu et al. (2019a).

Examples of methods employing pixel-wise supervision and domain generalization include the DC-CDN network (Yu et al., 2021), which outputs a face depth map, and the IADG method (Zhou et al., 2023), which mitigates instance-specific features to avoid domain bias.

3.3 ACTIVE METHODS

As mentioned in Chapter 1, active methods depend on user interaction for liveness detection. They usually follow one of three guidelines: based on involuntary interaction, information injection, and based on challenge-response.

In the following sections, each of these guidelines is presented in more detail, along with examples of relevant works from the literature. It is important to note that this division serves an educational purpose, and a single method may intersect with multiple guidelines.

3.3.1 Involuntary interaction

Approaches based on involuntary interactions typically utilize natural physiological actions that occur unconsciously. Similar to passive methods, this type of approach is minimally or not at all intrusive and can often be tested on passive datasets.

Following this trend, Hernandez-Ortega et al. (2020) uses remote photoplethysmography (rPPG), a technique that involves analyzing video sequences to detect subtle color changes in the human skin, which reveal the presence of blood under the tissue, to estimate the subject's heart rate in the video. This heart rate information is then used as an auxiliary measure to verify if the presented face is real. In a similar strategy, Liu et al. (2018b) employs a CNN-RNN model to estimate face depth and rPPG signals with sequence-wise supervision. The estimated depth and rPPG are fused to verify liveness.

Yan et al. (2012) utilizes non-rigid motion cues found in genuine faces, such as eye blinking, combined with an analysis of movement consistency between the face and background, along with the evaluation of imaging quality defects introduced in the fake face reproduction. Singh e Arora (2017) and Singh e Arora (2018) evaluate not only the eye-blinking pattern but also the movements made with the subject's mouth to identify attacks.

3.3.2 Information injection

Methods based on information injection, as the name suggests, add specific information at the moment of media capture that interacts with the user. Typically, the inserted information is designed to behave one way when presented with a real face and entirely differently when interacting with a fake face.

Adopting this premise, Farrukh et al. (2020) presents a sequence of light patterns on portions of the device screen. The reflections of the emitted light are then used to calculate surface normal vectors, estimating a 3D model of the surface. This model is used to differentiate a real face from 2D attacks. Additionally, the position of the flash serves as a signature, making the model more robust against injection and replay attacks.

Similarly, Zhang et al. (2021) adopts the idea of light pattern emission and use a CNN for depth map recovery and liveness classification, along with a regression branch for light CAPTCHA checking to search for the injected pattern in the user's face and eyes.

A multimodal FAS method for mobile devices is detailed in Kong et al. (2024). This approach utilizes three built-in sensors to capture raw data from both genuine and spoofed faces. The front camera captures RGB images, while the speaker emits a customized high-frequency signal and the microphone collects the reflected acoustic signal. The acoustic fingerprint is extracted from the audio data and converted into a spectrogram map. These features, along with the processed images, are then fed into a cross-modal fusion model to make the final decision.

3.3.3 Challenge-response

In systems that require user cooperation, also known as challenge-response systems, the user is instructed to perform a series of simple actions. The idea behind this type of approach is to place the Presentation Attack Instrument (PAI) in situations that highlight characteristic features of spoofs. It is intuitive to deduce that methods based on user interaction tend to compromise system usability in favor of the security of the liveness verification process. This tendency occurs due to the potential difficulty in recording a valid input, that is, the user's difficulty in understanding and executing the presented commands at the moment of input media acquisition.

The nature of the challenges presented to the user varies greatly. For example, in Sluganovic et al. (2016), the user must follow a pattern shown on the device screen with their eyes. The input video is then used to assess whether the user actually followed the presented pattern or completed the challenge suspiciously. A similar strategy is adopted in Shen et al. (2018), where a study was conducted addressing random iris movements, allowing for refinement of the displayed patterns and improving the performance of its predecessor. As discussed in these works, approaches based on following patterns with the eyes tend to face usability issues, as the act of following a pattern on the screen requires attention, and even a small deviation in gaze can compromise the authentication process. Furthermore, the exclusive use of gaze tracking may make the model more vulnerable to mask attacks or other more sophisticated attacks that leave the eye region exposed, allowing the attacker to solve the challenge.

Another example of a challenge-response-based approach is to present a random sequence of words, numbers, or syllables and ask the user to read them aloud. In McShane e Stewart (2017), this type of challenge is adopted, where the acquired audio is initially transcribed to text, and the same is done using lip-reading techniques from the video alone. It is then verified whether both media actually produced the word to be pronounced. Uzun et al. (2018) and Chou (2021) also use this challenge, but their methods assign a score representing the coherence between the phoneme captured in the audio and the video segment in which the audio appears, enhancing the system's robustness against more sophisticated attacks. However, voice-based approaches naturally have two disadvantages: they require capturing audio information, thus necessitating additional hardware infrastructure compared to other face spoof detection approaches. As mentioned in the cited works, another drawback is the interference from external environments, limiting the practical use of this approach to locations with minimal noise and good sound isolation.

In Ezz et al. (2023), the user is asked to emulate a sequence of three randomly selected facial expressions. The model was trained to detect four facial states (joy, anger, sadness, and neutral) and classify each expression's image individually as live or spoof, then combine the results to produce a final prediction. This approach has a disadvantage compared to more traditional FAS methods due to the need for prior enrollment of a neutral face as a reference, which may not be feasible for certain applications.

3.4 CONCLUDING REMARKS

Given the current state of FAS research, the lack of available data for studying active FAS methods remains a significant challenge. Additionally, the limited studies in this area often fail to rigorously validate their proposed methods, either by not evaluating them in robust scenarios or by employing overly complex active interactions that compromise usability.

To address these challenges, this work introduces a new active liveness detection database featuring a close-up challenge-response interaction. This database consists of facial videos recorded by volunteer subjects using their own smartphones and generated spoof samples using selected faces from public datasets, featuring five different PAs. The proposed protocols are designed to evaluate methods in challenging scenarios, while the collection methodology ensures high variability in camera models, backgrounds, ethnicities, genders, and lighting conditions. Furthermore, we developed a hybrid approach inspired by previous models to effectively detect personification attacks.

Chapter 4

PROPOSAL

This chapter explains the proposed approach consisting of creating an active dataset entitled UFPR Close-Up and proposing an active hybrid method for spoof detection.

4.1 UFPR Close-Up Dataset

The creation of the dataset consisted of four steps: developing an application for collecting samples, collecting live samples, selecting targets for presentation attacks, and collecting spoof samples. Each of these steps will be detailed below along with the current proposed protocols.

Due to private financial support for the creation of this dataset and in compliance with data protection agreements, all spoof samples are currently available upon request at `https://web.inf.ufpr.br/vri/databases/ufpr-closeup/`, but live samples will be released starting in April 2027.

4.1.1 The Close-Up liveness challenge

As mentioned earlier, active approaches can use a sequence of commands that the user must execute to prove their liveness. The close-up approach was designed to be used in conjunction with a digital screen displaying the content captured by the camera, and the sequence of prompts is as follows:

1. **Distant Align**: In this step, a small ellipse appears in the center of the device's screen, indicating the area where the user's face should be positioned. The user must fit their face within this ellipse.

2. **Hold Distant Position**: Once aligned, the user must remain still for at least one second. If the alignment is lost, the challenge will revert to the previous step, requiring the user to realign their face within the distant area.

3. **Close Align**: In this step, a larger ellipse appears on the screen, indicating a new, closer area where the user's face should be positioned. The alignment process follows the same criteria as in the first step but it is based on the new, larger region shown on the screen. Notice that to position the presented face inside the new area, the user must reduce the distance between the presented face and the device, therefore, getting close to the recording camera.

4. **Hold Close Position**: Finally, the user must remain still for at least one second within this close-up region. If the face moves outside the designated region, the challenge returns to the previous step, requiring the user to realign within the closer area.

Instructions to align the user's face in the indicated region and to hold position are shown to the user during the media capture as exemplified in Figure 4.1. It can be seen that to complete the challenge, the user must fit it's face in both distant and close positions, in this specific order. Therefore, it is guaranteed that the recorded media will contain a face in two different distances (close and distant).

(a) (b) (c) (d)

Figure 4.1: 4.1(a) - Distant alignment step; 4.1(b) - Distant waiting step; 4.1(c) - Close alignment step; 4.1(d) - Close waiting step

4.1.2 Mobile app

This section is summarized in (Kamarowski et al., 2023). This previous work describes the app development process, the problem of lack of disclosure active datasets, and three active interactions: Close-up, Head movements, and Flash. Notice that only the former one is in the scope of this work. To ensure a close-up movement pattern among all samples, live or spoof, a mobile app for sample capturing was developed for Android and iOS devices. Exclusively for the collection of live samples, the application included a user registration system, requesting information such as self-identified gender (male or female), an optional age input, and a declaration of acceptance of the terms of use. These terms outlined the objectives of data collection, how the data would be used, and sought user consent for the use and publication of the collected information for academic purposes. After registration, users could record sessions completing three liveness challenges, including the close-up challenge. Once a user recorded a session, they were restricted from recording new sessions for a period of 12 hours. This measure was implemented to enhance variability in conditions such as lighting, clothing, environment, and other factors, increasing the dataset diversity.

The recordings were made using the device's front camera to obtain 20 frames per second videos with close-up movement. This value was chosen based on empirical tests from an initial sample, which revealed that a higher sampling frequency degraded the application's performance, compromising its usability.

During capture, each frame was checked in real time for face position using the Google ML Kit API for face detection method (Google, 2021). The smaller region, shown in steps 1 and 2 of the close-up challenge, was positioned at the center of the screen, with the ellipse's width set to half the screen width and its height set to 1.3 times the ellipse's width. For the larger region, shown in steps 3 and 4, the ellipse was also centered on the screen, but its width was 95% of the device screen width, while its height remained 1.3 times the ellipse's width. Figure 4.2 depicts the dimension of each used region.

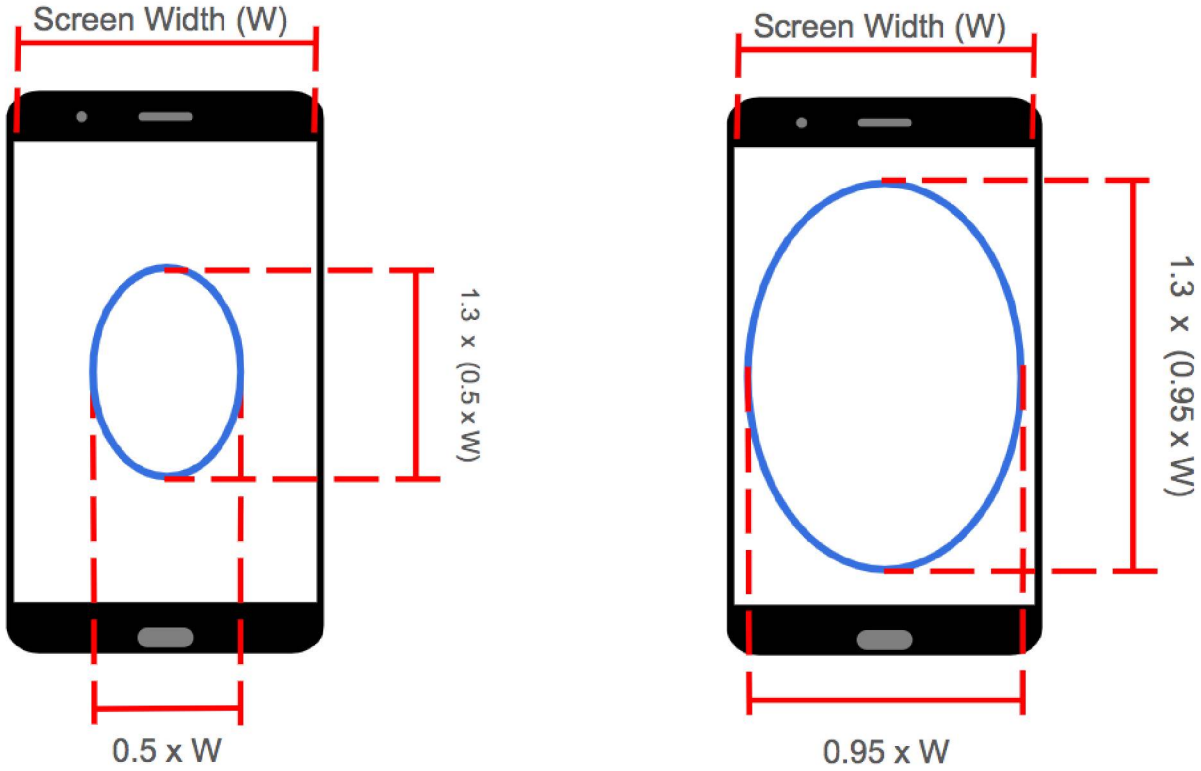


Figure 4.2: 4.2(a) - Dimensions of the distant region; 4.2(b) - Dimensions of the close region;

At every step, alignment was verified by calculating the Intersection over Union (IoU) between the detected face and the displayed region. In steps 1 and 3, alignment was considered correct if the IoU exceeded 0.8, whereas for steps 2 and 4, a lower threshold of 0.65 was used. This adjustment between adjacent steps was implemented to prevent minor oscillations while moving the face from causing the challenge to unexpectedly advance to the next step and then revert, a common issue when the IoU value is too close to the set threshold. This whole processing is summarized in Figure 4.3, the squares are the current state of the solving process of the close-up challenge and the diamond-shaped elements are checks done after each new frame.

In addition, the application automatically deletes recorded videos that do not successfully complete all four steps. Correct samples were uploaded to a private server to be stored.

4.1.3 Live samples acquisition

All live subjects are volunteers who agreed to participate in the data collection. To simulate a realistic use scenario, volunteers received no additional instructions on how to perform the close-up challenge. Furthermore, each volunteer used their personal device to capture live samples, resulting in a diverse array of camera types and qualities in the dataset.

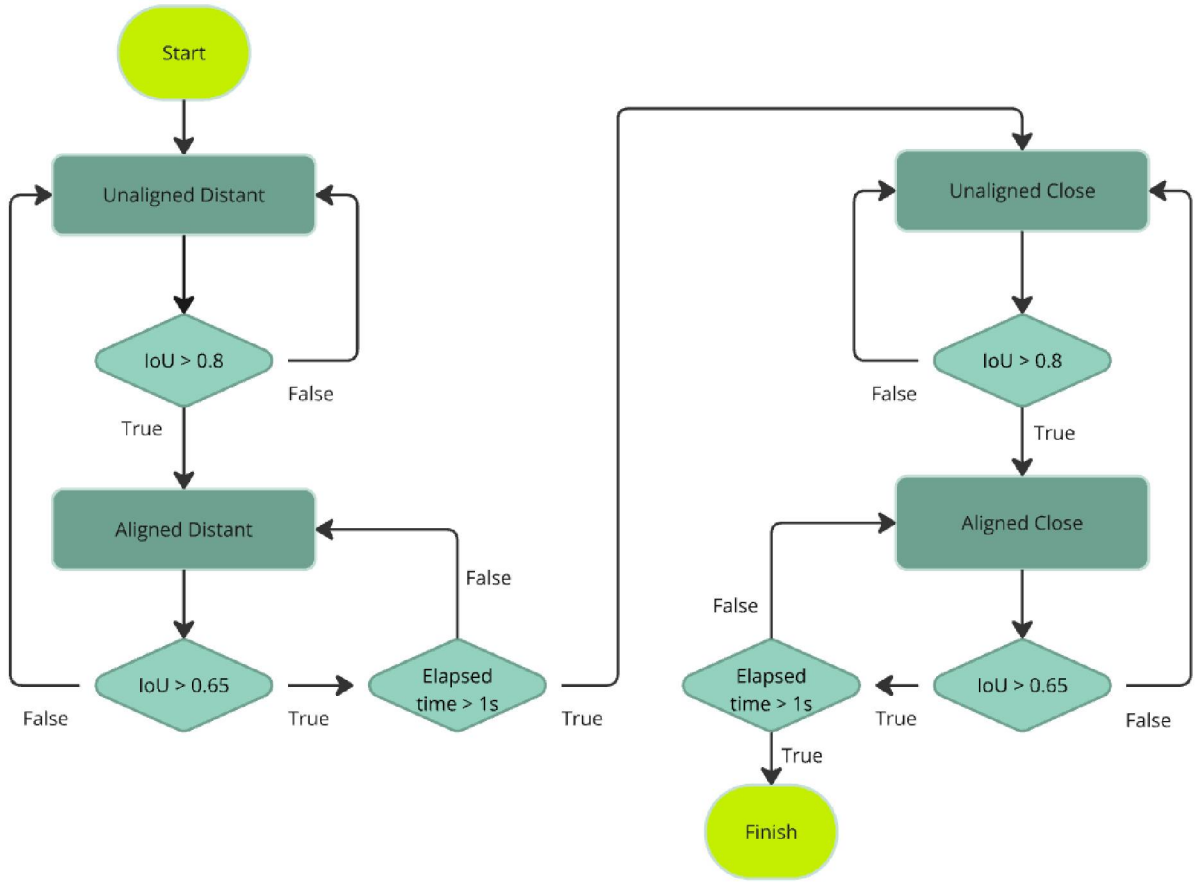


Figure 4.3: Fluxogram of the app's implementation of the close-up challenge

Participants were asked to record two sessions, but not limited to this amount, with a minimum interval of 12 hours between sessions. This approach ensured that the dataset included multiple samples from the same subject under varying conditions, such as different lighting and backgrounds, enhancing the variability of the collected data. Figure 4.4 depicts frames from two sessions of some live samples.

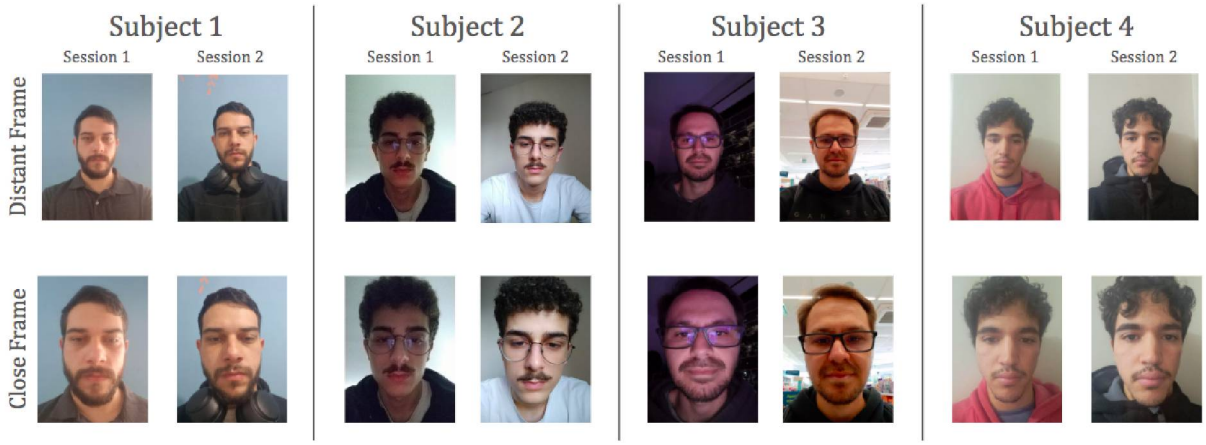


Figure 4.4: Close and Distant frames of four live subjects in different sessions.

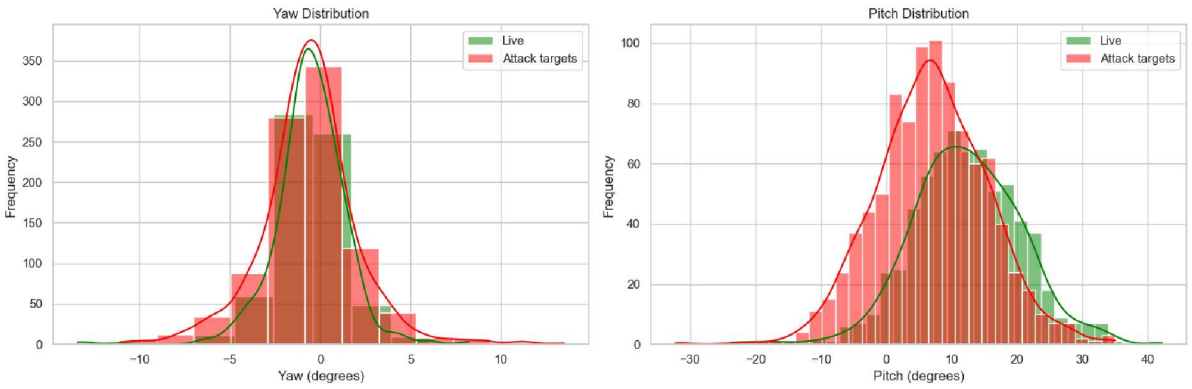


Figure 4.5: Final distribution of Yaw and Pitch angles

4.1.4 Target selection

In a real-world scenario, an attacker would likely not have access to media specifically created for facial liveness validation. Therefore, our approach diverges from most datasets in the literature by using publicly available images that an attacker might realistically access. These images, where the subjects are not necessarily engaged in facial authentication, may feature non-frontal postures and varied facial expressions. To generate spoof samples, we selected raw images from the CelebA (Liu et al., 2018c) dataset and videos from CelebV (Zhu et al., 2022).

To mitigate pose bias between live samples and attack targets, selected media were filtered considering the live pose distribution. This distribution was calculated by randomly selecting ten frames from each live sample, measuring yaw and pitch angles for each frame, and averaging the poses for each subject. In facial pose estimation, yaw and pitch are two key angles that describe the orientation of the head in 3D space. Yaw refers to the horizontal rotation of the head, indicating whether the face is turning left or right. Pitch, on the other hand, represents the vertical tilt of the head, showing whether the person is looking up or down. There is also a third angle for describing facial pose, roll, which represents the tilt to the side. This angle was not used as a criterion for face selection because, since applying an image rotation centered on the face, it is possible to modify the perceived roll angle. The head pose angles were estimated by detecting facial landmarks using MTCNN (Zhang et al., 2016), a widely adopted method known for its robustness in diverse scenarios. These landmarks were then mapped to a 3D face model using the PnP algorithm (Marchand et al., 2015) to accurately determine the head pose. This process calculates the face angles related to yaw, pitch, and roll movements from the extracted key points.

This produced pose distributions Yaw_{Live} and $Pitch_{Live}$. A relaxed interval for spoof targets was considered. That is, spoof target candidates outside the interval [-max($abs(Yaw_{Live})$),max(abs(Yaw_{Live}))] and [-max($abs(Pitch_{Live})$),max(abs($Pitch_{Live}$))] for target's media Yaw and Pitch angles, respectively were discarded. The remaining images were randomly selected maintaining the same gender distribution as the live samples. Figure 4.5 shows the final pose distributions for live and spoof subjects.

4.1.5 Spoof acquisition

The selected media were used as targets in presentation attacks solving the proposed active interaction prompt. To complete the close-up challenge, the attacker can move the device, keeping the attack instrument still, or do the opposite. Following the types of attacks documented

in the literature, our dataset includes some of the most commonly used techniques with a range of instruments for diverse spoofing attempts:

- **Photo**: The victim's image is printed on an A4 sheet of paper. This image can be large enough to show part of the victim's torso or focus only on the face. In the latter case, the photo is placed on the face of a mannequin to emulate a human body. The sheet of paper can be completely *flat* or *wrapped* around the mannequin's head, emulating the curvature of a human face.

- **Display**: Similar to the Photo attack, however, the image of the victim's face is shown on a digital screen. The material used in the attack can be a *television* or *monitor* screen (desktop or notebook).

- **Replay**: In contrast to a display attack, in a replay attack a video is shown on a digital screen, thus naturally capturing the victim's physiological aspects such as eye blinking, breathing, and head movements. Again, the materials used in the attack can be a *television* or *monitor* screen.

- **Scaling**: The idea of Scaling attacks is to generate a video that presents the same close-up movement pattern as a live sample. To generate these media, a random subset of the live samples was selected. And (i) the position of the center of the subject's face, (ii) the size of the image, and (iii) the axis of the eyes concerning the horizontal axis of the image were extracted from each frame. Then, for each live frame a new manipulated image with a selected target's face was created by using affine operations to rotate and translate the victim's face to match the positions of live frames, creating a video that follows the same pattern as the live sample. The resulting media was used to perform scaling attacks in the same way that the videos were used for replay.

- **Mask**: The masks were produced by printing on A4 sheets of paper photos of the faces of the attack targets on a real-life scale. These photos were then cut out in silhouette and shown next to a representation of the target's body. This representation could be a hanger with a piece of clothing on it or an object that adds volume and shape to a piece of clothing, such as a mannequin or the attacker's own body. In both cases, the mask is positioned in the region where the face would be located, resulting in mask attacks labeled as *hanger* and *cut-out*, respectively.

It is important to emphasize that no faces from the live media subset were used to generate spoof samples, meaning that the identities of subjects in the live subset are not present in any produced spoof media. Examples of frame samples picturing each PA and their PAI can be seen in Figure 4.6.

Figure 4.6: Examples of each Presentation Attack and their respective Presentation Attack Instruments.

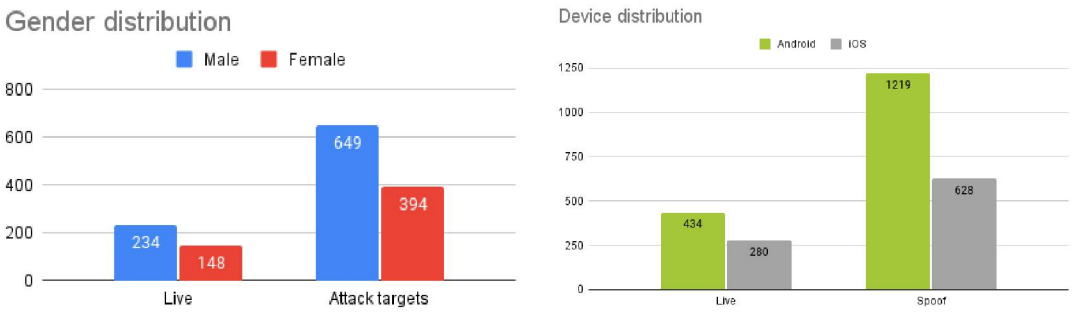
4.1.6 Dataset statistics

The distribution of dataset samples is summarized in Table 4.1. The first column contains the two possible classes (live or spoof), the second column displays the possible presentation attacks, and the third column lists the presentation attack instrument used on each PA. Finally, the last column details the number of samples per PAI on each PA.

Label	PA	PAI	#Samples
Spoof	Photo	Flat	186
		Wrapped	186
	Display	Monitor	186
		TV	186
	Replay	Monitor	186
		TV	186
	Scaling	Monitor	181
		TV	184
	Mask	Hanger	182
		Cut-out	184
Live	-	-	714

Table 4.1: Dataset sample distribution per Label, PA and PAI.

As previously mentioned, attack targets were selected based on the gender distribution of live samples. In addition, spoof samples were recorded according to device availability, but with an effort to maintain a similar proportion of samples captured on Android and iOS devices as in the live samples. Figure 4.7(a) details the final gender distribution, while Figure 4.7(b) shows the number of samples recorded using Android and iOS devices. It can be observed that the live distribution is approximately proportional to its counterpart across both figures.



(a) (b)

Figure 4.7: 4.7(a) - Distribution of male and female in both live subjects and selected attack targets; 4.7(b) - Distribution of samples recorded with Android and iOS in both live and spoof samples.

4.1.7 Protocols

We propose four protocols to evaluate the generalization capabilities of the algorithms for active PAD in scenarios with slight domain shifts on the UFPR-Close-up database. For all protocols, the dataset is divided into training, validation, and testing subsets with an approximate ratio of 3:1:1. The split maintains gender, number of different identities, and category of attack proportional among each subset. Additionally, each subject appears in only one subset, to avoid biased results due to memorization of identity facial cues. The proposed protocols are named as General, Unknown PAI, Unknown PA, and Unknown Device and defined as follows:

- **Protocol I - General**: The first protocol evaluates the ability to learn and generalize features in a scenario with almost no domain shift. All types of PA and PAI are present in the training, validation, and testing sets. This protocol is designed to obtain an overall result of the proposed PAD on the dataset and compare it with the results of the following protocols.

- **Protocol II - Unknown PAI**: Different PAI produces different noises after being captured by a digital camera. This protocol aims to evaluate the robustness of the proposed PAD against unseen PAI. Only attacks using Flat paper, Monitor, and Hanger masks are on the training set. The validation and testing set is composed of the remaining PAI (Wrapped paper, TV, and Cut-out). We follow the choice of PAI for training and validation/testing sets from other works in the literature (Boulkenafet et al., 2017b; Wang et al., 2023)

- **Protocol III - Unknown PA**: Following the same idea as the previous protocol, here a new sample arrangement is designed to study the ability to generalize learned features to unseen attacks. The training set has Photo, Display, and Scaling PAs, while Replay and Mask attacks appear only on validation and testing sets. Again, we follow the choice of PA from other works in the literature (Boulkenafet et al., 2017b; Wang et al., 2023)

- **Protocol IV - Unknown Device**: In the context of face anti-spoofing, it is crucial to include not only a wide representation of potential attacks under various lighting conditions and backgrounds but also a variety of cameras used for media acquisitions. In this protocol, samples are grouped based on the type of capturing device. Samples acquired with Android devices are used for training, whilst those obtained with iOS devices are assigned to the validation and testing sets.

Protocol	Subset	PA	PAI	Phone	#Live Videos	#Spoof Videos
Protocol I - General	Train	All	All	All	430	1073
	Val	All	All	All	147	376
	Test	All	All	All	137	398
Protocol II - Unknown PAI	Train	All	Flat Photo, Monitor, Hanger Mask	All	430	547
	Val	All	Wrapped Photo, TV, Cut-out Mask	All	147	196
	Test	All	Wrapped Photo, TV, Cut-out Mask	All	137	204
Protocol III - Unknown PA	Train	Photo, Display, Scaling	Flat photo, Wrapped Photo, Monitor and TV	All	430	641
	Val	Replay and Mask	Monitor, TV, Hanger Mask and Cut-out Mask	All	147	149
	Test	Replay and Mask	Monitor, TV, Hanger Mask and Cut-out Mask	All	137	157
Protocol IV - Unknown device	Train	All	All	Android	271	710
	Val	All	All	iOS	67	123
	Test	All	All	iOS	54	142

Table 4.2: Details of each protocol split.

The last three protocols exhibit a domain shift between the training subset and the validation and test subsets, following the same approach used in previous works (Boulkenafet et al., 2017b; Wang et al., 2023), i.e., defining a subset of features for training and another subset for validation and testing, therefore, emulating the ideal scenario where the validation and test sets follow similar distributions. Table 4.2 summarizes the proposed protocols and the average size of each partition.

4.2 HYBRID CLOSE-UP

The proposed method employs the close-up movement to capture facial features at various distances. This approach was inspired by the Camera Close-Up liveness detector (Castelblanco et al., 2022), sharing similarities in the creation of distortion feature vectors. The Hybrid Close-Up approach integrates texture embeddings with the concept of projective invariants, as defined by Riccio e Dugelay (2007), extracted from landmarks, with face embeddings in a fusion model.

The Hybrid Close-Up method is composed of three modules: Frame Selection, Feature Extraction, and Classification. Figure 4.8 provides an overview of the method, and each module is described in the followings.

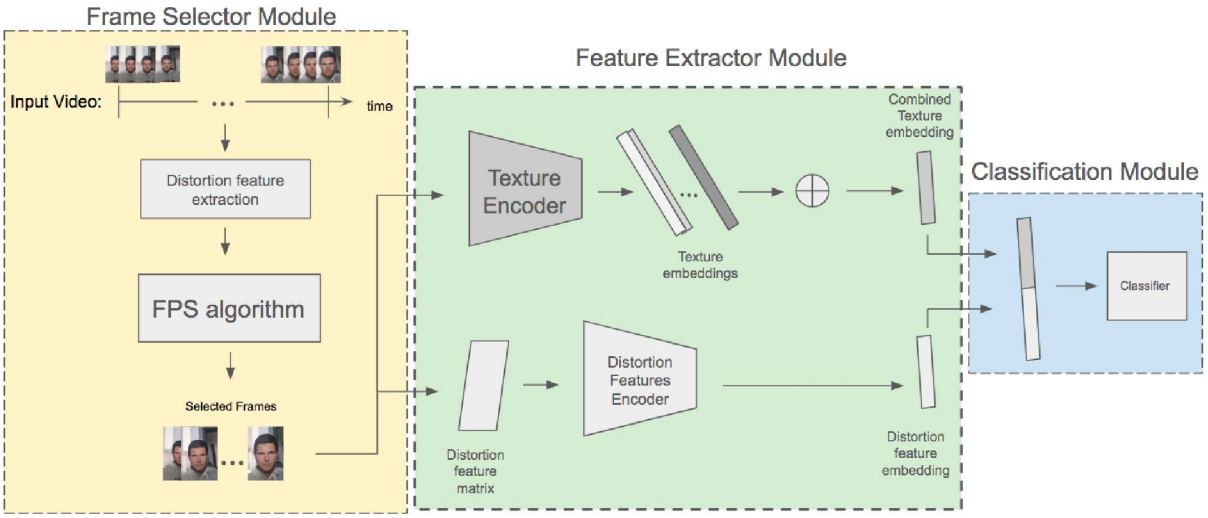


Figure 4.8: General pipeline used on Hybrid Close-up method.

4.2.1 Frame Selection

The first module of the pipeline is responsible for selecting the frames used in the next steps and this process is represented in the yellow box in Figure 4.8. In this work, we employ the MMOD CNN method (King, 2015) for face detection and an ensemble of regression trees described in Kazemi e Sullivan (2014) for landmark detection. Both methods were used by the mentioned baselines for similar tasks. First, the face detector is used to discard frames that do not contain a face. Next, from each valid frame the ladmarks of the shown face are extracted and the distortion features of the given face are computed. The distortion features are computed by calculating the Euclidean distances between all pairs of landmarks, excluding the distances between landmarks in the mouth region. Given a valid frame k, this process produces a distortion feature vector $d_k = (f_{k0}, f_{k1}, \ldots, f_{kM-1})$ of length M, where M is the number of distances between pairs of landmarks. The Frame Selection module also selects and computes the distortion feature vector of a special frame called reference frame, which is always the frame in the middle of the video.

Then, the computed distortion feature vectors are normalized. The normalization process of distortion features consists of the following: Given the distortion feature vector $d_k = (f_{k0}, f_{k1}, \ldots, f_{kM-1})$ of a selected frame-k and the distortion feature vector of the reference frame $d_{ref} = (f_{ref0}, f_{ref1}, \ldots, f_{refM-1})$. A normalized distortion feature vector d_{nk} is calculated as

$$d_{nk} = \left(\frac{f_{k0}}{f_{ref0}}, \ldots, \frac{f_{kM-1}}{f_{refM-1}} \right). \tag{4.1}$$

This process is done for each valid frame in the input video. The process of creating a normalized distortion feature vector from a selected frame and a reference frame is represented in Figure 4.9.

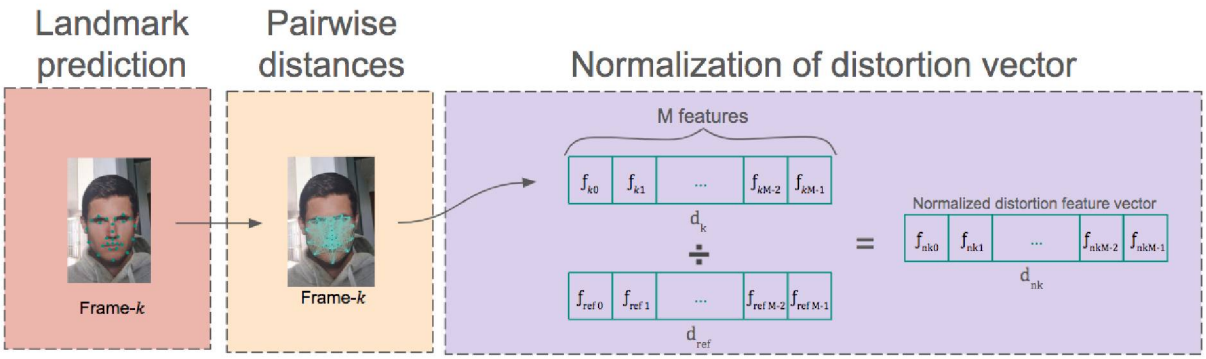


Figure 4.9: Representation of the process to extract distortion feature vectors from selected frames.

Next, N frames are selected by solving the Maximally Spread Subset Selection (MSSS) problem, defined as follows: Given a set of elements $P = \{p_0, p_1, ..., p_{V-1}\}$ of size V, a function $d(x, y)$ that measures the distance between a pair of elements $(x, y) \in P \times P$ and an integer N where $1 \leq N \leq |P|$. The expected output for an instance of the MSSS problem is a subset $P' \subseteq P$ with $|P'| = N$ where the sum of the distance between all pairs of elements in p' is maximal, or more precisely, ensures Equation 4.2, i.e.,

$$\max_{P' \subseteq P, |P'|=N} \sum_{\substack{x,y \in P' \\ x \neq y}} d(x, y) \tag{4.2}$$

The selection is made using the Farthest Point Sampling (FPS) algorithm (Eldar et al., 1997), which is a heuristic approach used to select a representative subset of points from a larger set and does not guarantee an optimal result for the MSSS problem. In our scenario, this algorithm uses the distortion feature vectors as the input set of points P, where each distortion vector is a point of dimension M. The well-known Euclidean distance was used as a distance function d. The FPS algorithm has a computational complexity of $O(VNM)$, where V is the total number of points. A generic implementation of this algorithm is shown in Algorithm 1. Note that the parameter N is defined in the experiments.

Algorithm 1 Farthest Point Sampling

Input: P (Set of points), N (Number of points to select)
Output: S (Subset of selected points)

Choose a random point $p_0 \in P$ and add to S
while $|S| < N$ **do**
    Find $p \in P \setminus S$ that maximizes $\min_{s \in S} d(p, s)$
    Add p to S
end while
return S

4.2.2 Feature Extraction

The next module is responsible for extracting texture features and processing the distortion feature vectors of selected frames. The distortion feature vectors are reorganized to form a matrix of distortion features $N \times M$, which is used as input to a distortion feature encoder, defined in Table 4.3, to extract the embeddings of the distortion features. The 1D Convolutions have kernel size 5, stride 1 and padding 1 while the 1D Max Pooling operations have kernel size 3, stride 2 and padding 1.

#Layer	Operations	input size	output size
1	Conv 1d	$N \times M$	$16 \times (M-2)$
	Batch normalization		
	ReLU		
2	Max Pooling 1d	$16 \times (M-2)$	$16 \times \frac{(M-1)}{2}$
3	Conv 1d	$16 \times \frac{(M-1)}{2}$	$32 \times \frac{(M-5)}{2}$
	Batch normalization		
	ReLU		
4	Max Pooling 1d	$32 \times \frac{(M-5)}{2}$	$32 \times \frac{(M-3)}{4}$
5	Conv 1d	$32 \times \frac{(M-3)}{4}$	$64 \times \frac{(M-11)}{4}$
	Batch normalization		
	ReLU		
6	Max Pooling 1d	$64 \times \frac{(M-11)}{4}$	$64 \times \frac{(M-7)}{8}$
7	Flatten	$64 \times \frac{(M-7)}{8}$	$8(M-7)$

Table 4.3: Distortion feature encoder architecture

The second type of feature summarizes the spatial information. These features are processed using texture encoders, which compute N texture embeddings of length E for each selected frame, excluding the reference frame. The embeddings are then merged using a spatial bottle neck structure based on learned maps defined in Table 4.4, producing a combined spatial embedding with the same length as the final distortion feature embedding. Variations of the texture encoder architecture are explored in Chapter 5.

#Layer	Operations	input size	output size
1	Fully connected	$N * E$	$8(M-7)$
	ReLU		

Table 4.4: Spatial bottle neck structure

4.2.3 Classification

The final module of the Hybrid Close-Up method is responsible for classifying the extracted embeddings as either live or spoof based on the computed features. This process begins by creating a final classification embedding, formed by concatenating the combined texture embedding with the distortion feature embedding, both of which are extracted in the preceding module. Then, this embedding is sent to a multi-layer perceptron (MLP) defined in Table 4.5 to predict the sample class.

#Layer	Operations	input size	output size
1	Fully connected	$1 \times 8(M-7) + 8(M-7)$	1×1024
	ReLU		
	Dropout (0.2)		
2	Fully connected	1×1024	1×512
	ReLU		
	Dropout (0.1)		
3	Fully connected	1×512	1×1
	Sigmoid		

Table 4.5: Description of MLP used for final classification

Chapter 5

EXPERIMENTS

This section discusses details regarding the used methodology and experimental results obtained exploring the proposed approach in different scenarios.

5.1 METHODOLOGY

The executed experiments and the steps for reproduction are presented. All experiments were conducted on a single machine equipped with the following GPUs: NVIDIA GeForce RTX 3060 and NVIDIA GeForce RTX 3090.

We followed the distortion feature extraction steps outlined in Castelblanco et al. (2022), utilizing the Python module of the dlib library (King, 2009) for face detection and landmark extraction across all implemented models. The used landmark extractor computes 68 points, with 10 of them being from the mouth region. Thus, the distortion feature vector of the proposed approach has a length of 2088 ($M = 2088$) (resulting from the distances of all pairs of points except all distances between pairs of the 10 points from the mouth region). Note that the Face Close-Up paper claim to use 66 landmarks but do not specify the used landmark extractor. Therefore, its distortion feature vector extraction routine was adapted to use 68 landmarks. This adaptation was trivially implemented using all pairwise distances between the extracted landmarks.

5.1.1 Dataset

All experiments used partitions of the proposed dataset detailed in Section 4.1. As mentioned earlier, no previous work on active approaches based on the close-up challenge discloses their used data. Therefore, the proposed dataset is the first publicly available dataset suitable for the proposed task.

5.1.2 Experiments

Initial experiments were conducted to define the ideal number of selected frames, texture encoder architecture and validate our proposed frame selection method. Followed by experiments featuring baseline methods and the proposed approach using the protocols (I to IV) of the dataset and finishing with an special test partition with spoof samples featuring live subjects. Results obtained using Protocol I were used as an upper bound comparison protocol, once it pictures the ideal scenario, with known PA and PAI. The final results on the testing subset were calculated using the weights and threshold that generated the lowest EER on the validation set.

Baseline methods were implemented based on descriptions available in their respective papers using PyTorch and other parameters (e.g., number of selected frames) are reproduced according to the original works. Both methods were trained using the ADAM optimizer, with a learning rate of 0.001, a batch size of 20, and using early stopping with the patience value set as 30 epochs[1]. Our proposed method was trained using the binary cross entropy loss function.

5.2 RESULTS

We now present a detailed analysis of the experimental results. First, we conduct a study to evaluate the impact of the number of selected frames. Then, we perform experiments to identify the optimal texture encoder for the Hybrid method. Next, we show the experiments comparing the Hybrid Close-up method operating with different frame selection approaches. Finally, we conclude with a comprehensive comparison of the baseline methods and the proposed approach across each protocol and tests with a special test set with different spoof samples.

5.2.1 Frame Selection Impact

As previously mentioned, the collected dataset has only two steps with strict temporal constraints, i.e., close and distant alignment with at least 1 second, each. With a sampling rate of 20 FPS, every dataset sample has at least 40 frames picturing faces. In this scenario, although the smallest step for frame selection is equal to 1, due to the elevated number of repetitions of each experiment, we adopted a step of 6. Based on these observations, we trained the proposed method with $N \in \{6, 12, 18, 24, 30, 36\}$. This experiment used only the proposed Protocol I, to get the expected upper bound of the proposed method.

Table 5.1 shows the results of this experiment. It can be seen that the ACER reaches its lowest value of 4.33% with 12 frames. After this value, a degradation in the method's performance is observed as the number of selected frames increases. In light of the presented results, the following experiments using the Hybrid Close-up method were conducted selecting 12 frames.

N	Validation		Test							
	EER (%)	EER th	Photo	Display	Scaling	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
6	2.10	0.870	0.72	0.98	1.84	0.00	5.08	5.08	5.24	5.16
12	**2.04**	**0.862**	**0.72**	**1.23**	**1.05**	**0.00**	**4.41**	**4.41**	**4.26**	**4.33**
18	2.40	0.937	0.60	1.23	1.58	0.00	5.12	5.12	5.60	5.36
24	2.97	0.846	0.84	1.85	1.71	0.00	7.12	7.12	5.21	6.16
30	2.68	0.710	0.72	1.60	1.32	0.00	7.24	7.24	5.56	6.39
36	2.59	0.779	0.84	1.75	1.58	0.02	7.03	7.03	6.25	6.64

Table 5.1: Experimental results on frame selection.

5.2.2 Encoder Impact

To extract spatial embeddings, we chose to use well-known encoder architectures designed for pattern recognition tasks. To do that, we used only the encoder portion of the networks ResNet-18, ResNet-34, ResNet-50 and ResNet-100. Note that ViT encoders are not employed here due to the reduced number of approaches adopting this technique in FAS tasks, especially in active scenarios. ViT models typically require large datasets to achieve optimal

[1]All implementations, including our proposal, are available at `https://github.com/BrunoHKC/Close_Up_Methods`.

performance, which is not feasible with the collected dataset containing only 1.503 video samples for training in Protocol I.

Table 5.2 shows the results of this experiment. The configuration that achieved the lowest errors (APCER, BPCER, and ACER) was the one using the ResNet-50 encoder, outperforming ResNet-18 and ResNet-34 by a wide margin. We hypothesize that ResNet-50's superior performance is due to its greater model capacity, enabling it to learn more complex features and capture finer details in the embeddings. Following this reasoning, we also suggest that ResNet-101 did not outperform ResNet-50 because the latter offers a complexity better suited to the dataset's characteristics, whereas ResNet-101 may have been more prone to overfitting. In light of the presented results, the following experiments using the Hybrid Close-up method were conducted using the ResNet-50 encoder.

Encoder	Validation		Test							
	EER (%)	EER th	Photo	Display	Scaling	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
ResNet-18	3.33	0.723	0.36	1.60	1.45	1.75	6.74	6.74	4.89	5.81
ResNet-34	2.24	0.714	0.24	1.11	1.97	0.27	5.90	5.90	5.54	5.72
ResNet-50	**2.04**	**0.86**	**0.72**	**1.23**	**1.05**	**0.00**	**4.41**	**4.41**	**4.26**	**4.33**
ResNet-101	2.01	0.871	1.44	1.85	1.32	0.00	4.85	4.85	4.21	4.53

Table 5.2: Experimental results on texture encoder.

5.2.3 Validation of the frame selection approach

To investigate if the proposed approach for frame selection is actually improving the Hybrid Close-Up performance, we evaluted the model using the proposed frame selection criteria, entitled as FPS-based in this experiment, using the approach used by the Face Close-Up, named Face Size - based and lastly, using the approach described in the Camera Close-Up's work, entitled Bin-based.

Table 5.3 shows the results of this experiment. It can be observed that the configuration achieving the lowest errors (APCER, BPCER, and ACER) was using the FPS-based approach, followed by the Bin-based and the face size-based approaches. We hypothesize that the improved performance of FPS-based frame selection stems from its sampling strategy, which prioritizes selecting the most dissimilar frames. This approach reduces redundancy and ensures that the proposed model receives more diverse and informative data. As expected, the Bin-based approach achieved similar results. The underlying reason is comparable—by selecting a fixed number of frames from each segment of a video, it promotes diversity across different portions of the input. However, it has a higher tendency to sample similar frames.

Furthermore, as discussed in Camera Close-Up's work, selecting frames based on face size may not be a generalizable approach. This method relies on a uniform movement pattern across all samples to ensure that each video contains frames representing the desired face sizes. However, in less constrained scenarios, such uniformity is unlikely, limiting the effectiveness of this technique.

In light of the presented results, we can observe that the proposed fame selection approach is more efficient in sampling relevant data. Therefore it validates the usage of the FPS-based frame selection method.

Frame Selection criteria	Validation		Test							
	EER (%)	EER th	Photo	Display	Scaling	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
Face size-based	2.31	0.931	0.24	1.47	2.36	0.81	6.66	6.66	4.43	5.55
Bin-based	2.36	0.762	0.60	1.35	1.58	0.00	4.55	4.55	4.70	4.62
FPS-based	**2.04**	**0.862**	**0.72**	**1.23**	**1.05**	**0.00**	**4.41**	**4.41**	**4.26**	**4.33**

Table 5.3: Experimental results on texture encoder.

5.2.4 Protocols

Finally, results using the baselines and the Hybrid Close-up method on each proposed protocol are described below. Experiments using the first protocol are shown in Table 5.4. It can be seen that the proposed method outperforms the baselines by a significant margin, scoring a lower BPCER, APCER and ACER for all PA. Several factors contribute to these results, but we hypothesize that the most significant one is the use of the full image as input to the model. This approach allows the model to leverage spatial cues beyond the face region, enhancing its ability to detect spoofs more effectively. As previously mentioned, values obtained using protocol I tend to be an upper bound once the training, validation, and testing set contains all types of PA, PAI, and devices.

Method	Validation		Test							
	EER (%)	EER th	Photo	Display	Scaling	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
Camera Close-Up	7.39	0.285	9.40	4.44	4.87	6.22	5.18	9.40	9.52	9.46
Face Close-Up	9.60	0.335	10.84	5.68	5.26	8.51	5.54	10.84	13.72	12.28
Hybrid Close-Up	**2.04**	**0.862**	**0.72**	**1.23**	**1.05**	**0.00**	**4.41**	**4.41**	**4.26**	**4.33**

Table 5.4: Experimental results using Protocol I.

In real-world scenarios, a spoof detection model will likely encounter spoof attacks created using PAIs that were not present in its training set. Given this challenge, generalization is a crucial property for any FAS.

To evaluate the model's ability to generalize to unknown PAIs, we conduct experiments under Protocol II, with the results presented in Table 5.5. It can be observed that the ACER of all methods increased compared to the values in Table 5.4, emphasizing the greater difficulty of this protocol. This effect is particularly evident in the case of Photo attacks, where the performance degradation is more pronounced. This increase was the primary factor raising the overall ACER from 9.46% to 12.54% for the Camera Close-Up method, from 12.28% to 20.59% for Face Close-Up, and from 4.33% to 9.13% for Hybrid Close-Up. The reason for this may be the differences among Photo Attack PAIs. Curved images emulate depth and exhibit 3D features that are absent in the flat images used for training. When confronted with these unfamiliar depth patterns in spoof samples from curved images, the model struggles to generalize effectively due to the significant deviation from its learned features. In contrast, other PAIs do not introduce such a pronounced domain shift, resulting in a less severe performance degradation.

Method	Validation		Test							
	EER (%)	EER th	Photo	Display	Scaling	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
Camera Close-Up	7.93	0.619	15.71	3.90	6.83	5.79	4.76	15.71	9.37	12.54
Face Close-Up	11.33	0.511	25.71	5.61	6.10	12.89	6.90	25.71	15.47	20.59
Hybrid Close-Up	**3.62**	**0.807**	**9.76**	**4.31**	**3.25**	**0.00**	**5.79**	**9.76**	**8.51**	**9.13**

Table 5.5: Experimental results using Protocol II

To study the effectiveness of proposed PAD methods against unseen PA, and consequently, some PAIs, protocol III isolates two out of five PAs. The obtained values for each of these

scenarios are described in Table 5.6. As we can see, all methods experienced performance degradation when handling Replay and Mask attacks, compared to their results for these attacks in Protocol I. Moreover, based on the ACER values obtained, we observe that performance against unknown PAIs (Protocol II) is nearly as challenging as the scenario explored in Protocol III.

Method	Validation		Test				
	EER (%)	EER th	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
Camera Close-Up	11.61	0.347	14.19	2.89	14.19	11.83	13.01
Face Close-Up	14.56	0.420	19.32	5.30	19.32	15.99	17.65
Hybrid Close-Up	**3.63**	**0.809**	**0.72**	**13.25**	**13.25**	**7.30**	**10.28**

Table 5.6: Experimental results using Protocol III.

Lastly, to verify the impact of the used acquisition camera, protocol IV trains proposed models on the Android-collected samples and evaluates their performance on the iOS-collected inputs. The obtained values for these scenarios are depicted in Table 5.7. Here, we see that although Photo, Replay, and Mask attacks were slightly easier to detect, the remaining PAs were more difficult to classify. This reveals that variations in capture sensors represent another significant concern for face PAD.

Method	Validation		Test							
	EER (%)	EER th	Photo	Display	Scaling	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
Camera Close-Up	5.40	0.317	2.97	9.41	5.77	5.38	3.33	9.41	11.73	10.57
Face Close-Up	8.61	0.301	17.03	6.47	11.92	5.38	10.56	17.03	17.22	17.12
Hybrid Close-Up	**2.99**	**0.843**	**0.63**	**1.07**	**2.57**	**0.00**	**5.56**	**5.56**	**11.11**	**8.33**

Table 5.7: Experimental results using Protocol IV.

Based on the experiments shown, we can see that the dataset has a high ACER in all methods across various scenarios. Therefore, the collected dataset is still challenging for the latest published active PAD. Additionally, the proposed method outperforms the baselines by a wide margin across all protocols, demonstrating its superior ability to generalize learned features in small domain shifts.

5.2.5 Test with special spoof samples

A modified test set was created in which spoof samples were generated using the live identities already present in the original test set. In other words, in this variation, the same identity appears in both the live and spoof classes. For this subset, one random frame from each live subject in the original test set was selected and used to generate a spoof sample for each PA. The only exception was the Replay attacks, where the entire video was utilized. The recording methodology for the new spoof samples followed the procedure outlined in the previous chapter. Table 5.8 presents the distribution of this new test partition.

Label	PA	PAI	#Samples
Spoof	Photo	Flat	36
		Wrapped	37
	Display	Monitor	36
		TV	37
	Replay	Monitor	36
		TV	37
	Scaling	Monitor	36
		TV	37
	Mask	Hanger	36
		Cut-out	37
Live	-	-	137

Table 5.8: Special test partition distribution per Label, PA and PAI.

It is hypothesized that using the same identities in both the live and spoof sets may degrade model performance. This hypothesis is based on the observation that using live images to generate spoofs makes the feature distributions of both classes more similar, as the images themselves are more alike. As a result, the decision frontier tends to be stricter and more complex, making it more difficult to separate the two classes. Moreover, using the same subjects in instances live and spoof samples ensures that both classes share the same facial feature distributions, preventing the model from leveraging identity-related cues to distinguish real faces from spoofs. It is important to emphasize that due to the sensitivity of the data used in this new test set, no spoof sample generated from a live identity can be disclosed.

Next, Table 5.9 presents the results of an experiment in which the Hybrid Close-Up method was trained and validated using the standard train and validation subset of the UFPR-Close-Up dataset, respectively, but tested with the modified test set described in this section.

Method	Validation		Test							
	EER (%)	EER th	Photo	Display	Scaling	Replay	Mask	Overall		
			APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	APCER (%)	BPCER (%)	ACER (%)
Camera Close-Up	7.28	0.278	11.48	5.87	6.35	9.84	7.41	11.48	9.39	10.43
Face Close-Up	9.66	0.393	12.61	6.06	6.39	11.38	7.69	12.61	13.64	13.13
Hybrid Close-Up	**2.10**	**0.879**	**3.04**	**4.61**	**3.94**	**8.76**	**6.95**	**8.76**	**4.31**	**6.54**

Table 5.9: Experimental results using test set with spoof samples generated from live midias.

The results show a substantial drop in performance when comparing the values in Table 5.4, changing the ACER metric from 4.33% to 6.54% for the Hybrid Close-Up case, supporting the proposed hypothesis.Along with a significant increase in ACER, a particularly notable pattern is the sharp rise of 8.76 percentage points in APCER for Replay attacks. We hypothesized that this can be explained by the nature of the spoof media used in standard Replay attacks, which often come from datasets capturing individuals in diverse activities. Consequently, these media may lack key characteristics present in the Close-Up authentication process, such as faces captured within the expected distance range. The same reasoning applies to the other PAs. That is, the images selected from CelebA to create the other PAs were captured in various contexts and may exhibit characteristics such as lighting conditions, a higher frequency of individuals wearing makeup, facial expressions, and other traits that are uncommon in live samples. This could potentially introduce a bias that the model may have learned for classification.

When live media are used to generate spoof samples, the live and spoof sets share the same identities. As a result, the inter-class distance is reduced, making the classification task significantly more challenging.

Chapter 6

CONCLUSION

A comprehensive overview of the state of the art in face presentation attack detection was presented in this work. Prominent datasets were also listed to highlight the challenges within the active face anti-spoofing field, particularly the scarcity of publicly available data suitable for developing and evaluating solutions, which are essential for ensuring reproducibility and fostering collaborative research.

To address this issue, we focus on creating a new FAS active dataset based on close-up interactions, contributing to the field by establishing a common dataset and evaluation protocols, paving the way for further academic advancements.

The conducted experiments demonstrate that the collected dataset poses significant challenges to recent active close-up methods proposed in the literature. While the proposed method outperforms existing approaches in the evaluated scenarios being a strong evidence in favor of our hypothesis that the spacial and texture features extracted from image embeddings complement the distortion features, improving the model's performance. However, the proposed method still struggles with generalization issues, as observed in Protocols II, III, and IV. Moreover, as show in Section 5.2.5, we observe that the usage of the same identities in both live and spoof samples may be a challenging scenario.

Future research may explore several avenues for improvement. One promising direction is refining the frame selection process by incorporating image quality metrics to ensure more informative samples. Another potential enhancement involves integrating strategies commonly used in passive PAD methods within the classification module, such as auxiliary tasks like depth map estimation.

Furthermore, future studies could investigate alternative active interactions, such as the head movement challenge-response method described in (Castelblanco et al., 2022), where users rotate their heads toward prompted positions. Another promising approach is the flash challenge, which projects different light patterns onto the user's face to reveal live cues like depth and shadow projections, as utilized in Liu et al. (2019b). Both challenges are contemplated by the UFPR dataset.

REFERENCES

Agarwal, A., Yadav, D., Kohli, N., Singh, R., Vatsa, M. e Noore, A. (2017). Face presentation attack with latex masks in multispectral videos. Em *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Anjos, A. e Marcel, S. (2011). Countermeasures to photo attacks in face recognition: A public database and a baseline. Em *2011 International Joint Conference on Biometrics (IJCB)*, páginas 1–7.

Antil, A. e Dhiman, C. (2025). Unmasking deception: A comprehensive survey on the evolution of face anti-spoofing methods. *Neurocomputing*, 617:128992.

Boulkenafet, Z., Komulainen, J. e Hadid, A. (2017a). Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145.

Boulkenafet, Z., Komulainen, J., Li, L., Feng, X. e Hadid, A. (2017b). OULU-NPU: A mobile face presentation attack database with real-world variations. Em *2017 12th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2017)*, páginas 612–618.

Castelblanco, A., Rivera, E., Solano, J., Tengana, L., López, C. e Ochoa, M. (2022). Dynamic face authentication systems: Deep learning verification for camera close-up and head rotation paradigms. *Comput. Secur.*, 115(C).

Chingovska, I., Anjos, A. e Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. Em *2012 BIOSIG - Int. Conf. of Biometrics Special Interest Group (BIOSIG)*, páginas 1–7.

Chou, C.-L. (2021). Presentation attack detection based on score level fusion and challenge-response technique. *The Journal of Supercomputing*, 77.

de Freitas Pereira, T., Anjos, A., De Martino, J. M. e Marcel, S. (2013). Lbp-top based countermeasure against face spoofing attacks. Em Park, J.-I. e Kim, J., editores, *Computer Vision - ACCV 2012 Workshops*, páginas 121–132, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eldar, Y., Lindenbaum, M., Porat, M. e Zeevi, Y. (1997). The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315.

Ezz, M., Mostafa, A. e Elshenawy, A. (2023). Challenge-response emotion authentication algorithm using modified horizontal deep learning. *Intell. Autom. Soft Comput.(IASC)*, 35:3659–3675.

Farrukh, H., Aburas, R. M., Cao, S. e Wang, H. (2020). Facerevelio: a face liveness detection system for smartphones with a single front camera. Em *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, MobiCom '20, New York, NY, USA. Association for Computing Machinery.

Garg, S., Mittal, S., Kumar, P. e Anant Athavale, V. (2020). DeBNet: Multilayer deep network for liveness detection in face recognition system. Em *2020 7th Int. Conf. on Signal Processing and Integrated Networks (SPIN)*, páginas 1136–1141.

George, A. e Marcel, S. (2021). On the effectiveness of vision transformers for zero-shot face anti-spoofing. Em *2021 IEEE International Joint Conference on Biometrics (IJCB)*, páginas 1–8.

Google (2021). Google ml kit: Flutter package. `https://pub.dev/packages/google_ml_kit`. Retrieved January 18, 2025.

Hernandez-Ortega, J., Tolosana, R., Fierrez, J. e Morales, A. (2020). Deepfakeson-phys: Deepfakes detection based on heart rate estimation.

ISO/IEC (2023). Information technology — biometric presentation attack detection — part 3: Testing and reporting. Standard ISO/IEC 30107-3:2023(E), International Organization for Standardization.

Jain, A. K., Ross, A. e Pankanti, S. (2006). Biometrics: a tool for information security. *IEEE transactions on information forensics and security*, 1(2):125–143.

Kamarowski, B., Almeida, R., Biesseck, B., Granada, R., Führ, G. e Menotti, D. (2023). Multi-challenge database for active liveness. Em *Anais Estendidos da XXXVI Conference on Graphics, Patterns and Images*, páginas 109–114, Porto Alegre, RS, Brasil. SBC.

Kamarowski, B., Almeida, R., Biesseck, B., Granada, R., Führ, G. e Menotti, D. (2024). Hybrid method for active face anti-spoofing based on close-up challenge. Em *Anais Estendidos da XXXVII Conference on Graphics, Patterns and Images*, Porto Alegre, RS, Brasil. SBC.

Kazemi, V. e Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 1867–1874.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

King, D. E. (2015). Max-margin object detection. *arXiv preprint arXiv:1502.00046*.

Komulainen, J., Hadid, A. e Pietikäinen, M. (2013). Context based face anti-spoofing. Em *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, páginas 1–8.

Kong, C., Zheng, K., Liu, Y., Wang, S., Rocha, A. e Li, H. (2024). M3fas: An accurate and robust multimodal mobile face anti-spoofing system. *IEEE Transactions on Dependable and Secure Computing*.

Kumar, S., Singh, S. e Kumar, J. (2017). A comparative study on face spoofing attacks. Em *2017 International Conference on Computing, Communication and Automation (ICCCA)*, páginas 1104–1108. IEEE.

Le, B. M. e Woo, S. S. (2024). Gradient alignment for cross-domain face anti-spoofing.

Li, H., Li, W., Cao, H., Wang, S., Huang, F. e Kot, A. C. (2018). Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809.

Li, Y., Wang, Z., Li, Y., Deng, R., Chen, B., Meng, W. e Li, H. (2019). A closer look tells more: A facial distortion based liveness detection for face authentication. Em *Asia CCS '19: ACM Asia Conference on Computer and Communications Security*, páginas 241–246.

Liu, W., Wei, X., Lei, T., Wang, X., Meng, H. e Nandi, A. K. (2021). Data fusion based two-stage cascade framework for multi-modality face anti-spoofing. *IEEE Transactions on Cognitive and Developmental Systems*, páginas 1–1.

Liu, Y., Jourabloo, A. e Liu, X. (2018a). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. Em *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Y., Jourabloo, A. e Liu, X. (2018b). Learning deep models for face anti-spoofing: Binary or auxiliary supervision.

Liu, Y., Stehouwer, J., Jourabloo, A. e Liu, X. (2019a). Deep tree learning for zero-shot face anti-spoofing. Em *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 4675–4684.

Liu, Y., Tai, Y., Li, J., Ding, S., Wang, C., Huang, F., Li, D., Qi, W. e Ji, R. (2019b). Aurora guard: Real-time face anti-spoofing via light reflection. *CoRR*, abs/1902.10311.

Liu, Z., Luo, P., Wang, X. e Tang, X. (2018c). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11.

Marcel, S., Nixon, M. S., Fierrez, J. e Evans, N. (2019). *Handbook of biometric anti-spoofing: Presentation attack detection*, volume 2. Springer.

Marchand, E., Uchiyama, H. e Spindler, F. (2015). Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651.

McShane, P. e Stewart, D. (2017). Challenge based visual speech recognition using deep learning. Em *2017 12th Int. Conf. for Internet Technology and Secured Transactions (ICITST)*, páginas 405–410.

Ming, Z., Visani, M., Luqman, M. M. e Burie, J.-C. (2020). A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices. *Journal of imaging*, 6(12):139.

Mostaani, Z., George, A., Heusch, G., Geissbühler, D. e Marcel, S. (2020). The high-quality wide multi-channel attack (HQ-WMCA) database. *CoRR*, abs/2009.09703.

Patel, K., Han, H. e Jain, A. K. (2016). Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283.

Quan, R., Wu, Y., Yu, X. e Yang, Y. (2021). Progressive transfer learning for face anti-spoofing. *IEEE Transactions on Image Processing*, 30:3946–3955.

Raheem, E. A., Ahmad, S. M. S. e Adnan, W. A. W. (2019). Insight on face liveness detection: A systematic literature review. *International Journal of Electrical & Computer Engineering (2088-8708)*, 9(6).

Riccio, D. e Dugelay, J.-L. (2007). Geometric invariants for 2d/3d face recognition. *Pattern Recognit. Lett.*, 28:1907–1914.

Rui, Z. e Yan, Z. (2018). A survey on biometric authentication: Toward secure and privacy-preserving identification. *IEEE access*, 7:5994–6009.

Shen, M., Liao, Z., Zhu, L., Mijumbi, R., Du, X. e Hu, J. (2018). Iritrack: Liveness detection using irises tracking for preventing face spoofing attacks.

Singh, M. e Arora, A. (2017). A robust anti-spoofing technique for face liveness detection with morphological operations. *Optik*, 139:347–354.

Singh, M. e Arora, A. (2018). A novel face liveness detection algorithm with multiple liveness indicators. *Wireless Personal Communications*, 100:1677–1687.

Sluganovic, I., Roeschlin, M., Rasmussen, K. e Martinovic, I. (2016). Using reflexive eye movements for fast challenge-response authentication. Em *CCS '16: ACM SIGSAC Conference on Computer and Communications Security*, páginas 1056–1067.

Sun, Y., Liu, Y., Liu, X., Li, Y. e Chu, W.-S. (2023). Rethinking domain generalization for face anti-spoofing: Separability and alignment.

Tan, X., Li, Y., Liu, J. e Jiang, L. (2010). Face liveness detection from a single image with sparse low rank bilinear discriminative model. Em Daniilidis, K., Maragos, P. e Paragios, N., editores, *Computer Vision – ECCV 2010*, páginas 504–517, Berlin, Heidelberg. Springer Berlin Heidelberg.

Uzun, E., Chung, S., Essa, I. e Lee, W. (2018). rtCaptcha: A real-time captcha based liveness detection system. Em *Conference: The Network and Distributed System Security Symposium (NDSS)*.

Wang, D., Guo, J., Shao, Q., He, H., Chen, Z., Xiao, C., Liu, A., Escalera, S., Escalante, H. J., Lei, Z. et al. (2023). Wild face anti-spoofing challenge 2023: Benchmark and results. Em *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, páginas 6379–6390.

Wang, J., Zhang, J., Bian, Y., Cai, Y., Wang, C. e Pu, S. (2021a). Self-domain adaptation for face anti-spoofing. *CoRR*, abs/2102.12129.

Wang, Y., Song, X., Xu, T., Feng, Z. e Wu, X.-J. (2021b). From RGB to depth: Domain transfer network for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:4280–4290.

Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T. e Wang, Z. (2022). Domain generalization via shuffled style assembly for face anti-spoofing.

Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F. e Lei, Z. (2020). Deep spatial gradient and temporal depth learning for face anti-spoofing. Em *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 5041–5050.

Wen, D., Han, H. e Jain, A. K. (2015). Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761.

Yan, J., Zhang, Z., Lei, Z., Yi, D. e Li, S. Z. (2012). Face liveness detection by exploring multiple scenic clues. *2012 12th Int. Conf. on Control Automation Robotics & Vision (ICARCV)*, páginas 188–193.

Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z. e Zhao, G. (2023). Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5609–5631.

Yu, Z., Qin, Y., Zhao, H., Li, X. e Zhao, G. (2021). Dual-cross central difference network for face anti-spoofing. *CoRR*, abs/2105.01290.

Yu, Z., Wan, J., Qin, Y., Li, X., Li, S. Z. e Zhao, G. (2020a). Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023.

Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F. e Zhao, G. (2020b). Searching central difference convolutional networks for face anti-spoofing. Em *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 5294–5304.

Zhang, J., Tai, Y., Yao, T., Meng, J., Ding, S., Wang, C., Li, J., Huang, F. e Ji, R. (2021). Aurora guard: Reliable face anti-spoofing via mobile lighting system.

Zhang, K., Zhang, Z., Li, Z. e Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Zhang, K.-Y., Yao, T., Zhang, J., Tai, Y., Ding, S., Li, J., Huang, F., Song, H. e Ma, L. (2020a). Face anti-spoofing via disentangled representation learning. Em *European Conference on Computer Vision*, páginas 641–657. Springer.

Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J. e Liu, Z. (2020b). CelebA-spoof: Large-scale face anti-spoofing dataset with rich annotations. Em *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, páginas 70–85. Springer.

Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D. e Li, S. Z. (2012). A face antispoofing database with diverse attacks. Em *2012 5th IAPR Int. Conf. on Biometrics (ICB)*, páginas 26–31.

Zheng, W., Yue, M., Zhao, S. e Liu, S. (2021). Attention-based spatial-temporal multi-scale network for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):296–307.

Zhou, Q., Zhang, K.-Y., Yao, T., Lu, X., Yi, R., Ding, S. e Ma, L. (2023). Instance-aware domain generalization for face anti-spoofing.

Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z. e Loy, C. C. (2022). Celebv-hq: A large-scale video facial attributes dataset. Em *European conference on computer vision*, páginas 650–667. Springer.