Hybrid method for active face anti-spoofing based on close-up challenge

Bruno Kamarowski*, Raul Almeida*, Bernardo Biesseck*, Roger Granada†, Luiz Coelho†, David Menotti*
* Federal University of Paraná, Curitiba, PR, Brazil {bhkc18, rgpalmeida, bernardo, menotti}@inf.ufpr.br
‡unico - idTech, Brazil {roger.granada, luiz.coelho}@unico.io

***Abstract*—Facial authentication on mobile devices has become prevalent in various applications. Face Liveness Detection, or Face Anti-Spoofing (FAS), focuses on identifying attempts by malicious users to impersonate someone else or hide their own identity. One specific branch within this field is active liveness detection, which involves analyzing both the input signal and user behavior while they perform a required task to verify the authenticity of the presented face. Despite the significant amount of research in FAS, active liveness detection remains mostly underexplored. This gap has led to outdated methods, insufficient testing of proposed active techniques in diverse scenarios, and a lack of comparative analysis between different approaches. In this paper, we explore these differences by comparing the performance of the latest existing close-up methods with baseline models using ResNet-18 and ResNet-50. Furthermore, we introduce a new model that builds on previous work, combining projective invariants with facial embedding for robust feature extraction. This approach directly improves upon existing techniques, surpassing other baselines in detecting spoofing attempts.**

I. Introduction

The implementation of facial recognition technologies in everyday systems has become a highly sensitive issue, even for individuals without expertise in the field. When such technology is employed for biometric authentication using images or videos, it is crucial to complement this verification with a liveliness check of the captured media. This check determines whether the presented face is indeed genuine or a malicious attempt to impersonate another individual using a counterfeit face.

Various facial liveliness attacks are documented in the literature, which can be categorized into injection attacks and presentation attacks [1]. Injection attacks occur when an attacker overwrites or bypasses the media acquisition process, injecting a custom file of their interest [2]. Presentation attacks, the most frequently discussed in the literature, involve showing a Presentation Attack Instrument (PAI) to the camera during media capture. The nature of presentation attacks is highly diverse, including the use of printed photos, digital screens, realistic synthetic masks, and numerous other techniques to impersonate a target or conceal the attacker's identity [3].

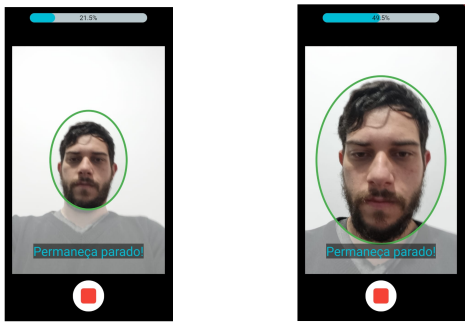
To mitigate such attacks, spoof detectors, also known as presentation attack detectors, are developed. These detectors are classified into two categories: passive and active [1]. Passive detectors do not require any special interaction during media acquisition. In contrast, active detectors rely on some form of user interaction to verify the presence of a legitimate person in front of the camera during authentication.

In the field of active liveness detection, a range of user interactions can be employed, including involuntary responses or physiological reflexes such as natural head or eye movements, blinking, or pupil dilation [4]–[6]. There are also approaches based on introducing stimuli during capture, such as sound or light patterns, and analyzing the response to determine if it interacts with a real face or a PAI [7]–[10]. Another technique involves asking the user to perform a simple task during media capture, such as smiling, nodding, or intentionally blinking [11], [12]. This approach often sacrifices system usability to enhance liveness verification. By making it more difficult to replicate a real face performing specific movements, it increases the challenge's difficulty and provides additional information about the dynamic aspects of the task, such as capturing the user's face from multiple angles during the movement.

Considering the tradeoff between usability and security in challenge-response-based approaches, in this work, we explore the close-up challenge, which involves two phases. In the initial phase, the user must position their face within a small area highlighted on the screen, maintaining a specific distance from the device. Once properly aligned and held in place for a brief period, the second phase is initiated. In this phase, the user is prompted to position their face within a larger area displayed on the screen, requiring them to move closer to the device. Figure 1 illustrates the two stages of this challenge. We believe that the close-up challenge is straightforward enough to minimally impact the usability of any system requiring liveness verification, yet it is capable of providing valuable information for accurate liveness detection.

Although the field of liveness detection is well-established and extensively discussed, the active liveness branch remains underdeveloped. One indicator of this is the scarcity of publicly available active datasets, often leading to new solutions being developed and evaluated with in-house data that cannot be shared due to its sensitive nature. Moreover, previous active liveness studies, such as Face Close-up [13] and Camera Close-up [14], have frequently been limited by datasets with too few subjects or controlled environments that do not accurately reflect real-world conditions.

To address these limitations, we have compiled a com-



(a) (b)

Fig. 1. 1a - Distant alignment step; 1a - Close alignment step.

prehensive dataset[1] comprising 683 live videos and 1027 spoof videos, captured in both indoor and outdoor settings without constraints and featuring over 1000 distinct identities. Additionally, we developed a new presentation attack detector called Hybrid Close-up, which outperforms two recent close-up movement-based approaches. We also compare its performance with two ResNet variants, emulating a simple passive approach.

The structure of this paper is as follows: Section II reviews previous works in the field of face liveness detection. Section III details the Hybrid Close-up method. Section IV presents the experimental results and outlines the reproduction steps. Finally, Section V summarizes the findings and concludes the paper with some discussions for future approaches.

II. Related Work

In this section, we present published works related to FAS. Section II-A discusses important passive liveness datasets, Section II-B summarizes studied passive face anti-spoofing methods, and Subsection II-C presents studied active face anti-spoofing methods.

A. Datasets for Passive Facial liveness

The advancement in the area of passive liveness detection has driven the proposal and creation of various datasets for this task. More recent studies consider not only the number of individuals and attacks used but also the variety of scenarios, lighting, camera quality, and diversity of the individuals that make up the dataset. Table I summarizes some of the most popular datasets in the literature and includes data from the datasets presented together with the Face Close-up and Camera Close-up methods [13], [14], as well as our dataset collected specifically for this study.

It is important to note that the cited passive datasets are not suitable for active approaches with the close-up paradigm, as they do not include any kind of interaction of this nature. An exception to this statement is the SiW dataset [23], which includes a partition of images with individuals performing non-trivial tasks to assess the robustness of passive methods to variations in pose and distance. Moreover, the studies mentioned in Section II-C did not disclose the data used in their respective experiments. Therefore, to the best of our knowledge, there are no publicly available datasets designed for the active liveness detection scenario using the close-up paradigm.

[1]In this work, the dataset is briefly presented; its complete version will be described in a future journal paper.

B. Passive Facial liveness Mechanisms

Over the years of liveness studies, approaches for Face Anti-Spoofing (FAS) have transitioned from detecting simple handcrafted features to learning feature maps. Among the earlier strategies, Boulkenafet et al. [29] described facial appearance by applying Fisher vector encoding to features extracted from different color spaces for FAS. Chingovska et al. studied the effectiveness of using texture features based on Local Binary Patterns and their variations in classification.

More recent methods may use handcrafted features in a hybrid manner, combining them with features extracted from a deep neural network [30], [31]. Traditional Deep Learning-based methods use end-to-end CNNs to learn a direct mapping from face image to liveness label, relying on direct [32] or pixelwise [33], [34] supervision for model training.

Generalized Deep Learning methods take a step further and aim to be robust against characteristic changes (for example, variations in the input sensor or in attack types). Methods may focus on domain generalization (when a model is trained only once) [35], [36] or adaptation (when the model undergoes an adaptation algorithm that leverages test data) [37], [38], or generalizing to unseen attack types with zero- and few-shot strategies (where no or only a small quantity of data is provided for training) as well as anomaly detection (where the model learns an accurate representation for live samples instead of learning the characteristics of spoofs) [24]. Examples of pixel-wise supervision and domain generalization include the DC-CDN network [33], which produces a face depth map as output, and the IADG method [35], which whitens instance-specific features to avoid domain bias.

C. Active Facial Liveness Mechanisms

As mentioned earlier, active methods depend on user interaction for liveness detection. They can be classified into three main lines: based on involuntary interaction, based on voluntary interaction, and injected information.

Face anti-spoofing based on involuntary interaction typically employs features from natural physiological movements. Some works extract specific features [6], [39], classifying a sample as real or spoof based on blinking patterns and lip movement patterns. Pupil movement is also used as a cue for liveness detection [5], and it has been experimented with combining more of such cues [4] (namely blinking, mouth movements, face-background consistency, and other aspects of samples) for facial liveness detection.

In strategies based on injected data, additional information is introduced during media capture. The usage of light pattern emissions has been studied [7], as well as using a CNN

TABLE I
STUDIED DATASETS' MAIN CHARACTERISTICS.

Dataset	Samples	Subjects	Attack types	User interaction
NUAA [15]	5105 real, 7509 spoof	15	1	Passive
PRINT-ATTACK [16]	200 real, 200 spoof	50	1	Passive
CASIA [17]	150 real, 450 spoof	50	3	Passive
Replay-Attack [18]	200 real, 1000 spoof	50	3	Passive
MSU-MFSD [19]	110 real, 330 spoof	55	3	Passive
MSU-USSA [20]	1140 real, 9120 spoof	1140	2	Passive
MLFP [21]	150 real, 1200 spoof	10	2	Passive
Oulu-NPU [22]	990 real, 3960 spoof	55	4	Passive
SiW [23]	1320 real, 3300 spoof	165	6	Multiple angles, face expressions and the subjects move
SiW-M [24]	660 real, 968 spoof	493	13	Passive
HQ-WMCA [25]	555 real, 2349 spoof	51	10	Passive
DMAD [26]	900 real, 1800 spoof	300	6	Passive
Celeb A-Spoof [27]	156,384 real, 469,153 spoof	10,177	6	Passive
WFAS [28]	529,571 real, 853,729 spoof	469,920	18	Passive
Face Close-up dataset [13]	710 real videos, 4970 spoof videos	71	3	Close up
Camera Close-up dataset [14]	89 real videos, 2537 spoof videos	41	5	Close up
Ours	683 real videos, 1027 spoof videos	372 live, 709 spoof	5	Close up

for depth map recovery and liveness classification with a regression branch that performs light CAPTCHA checking to search for the injected pattern in the user's face and eyes [8]. Another work focuses on emitting sound signals while the user is engaged in a simple task, analyzing the recovered signal (i.e., the echo of the emitted signal) to extract 3D facial geometry properties, and feeding these properties to an SVM classifier [10].

In systems relying on user cooperation, also known as challenge-response systems, the user is instructed to perform simple actions. For instance, the user might be required to follow a displayed pattern with their eyes or to point their eyes at a designated point on the screen [40], [41]. If the user fails the challenge or completes it with suspicious patterns, they are classified as spoofed. Another example is to ask the user to pronounce a randomized sequence of words and detect liveness based on the consistency between mouth and face movement and the audio sample [42]–[44]. It is also possible to verify facial three-dimensionality through projective invariants from a sequence of head movements [45].

Regarding the close-up paradigm, the Face Close-Up method [13] selects a reference frame along with a set of frames from an input video based on the face size relative to the entire image. It computes facial landmarks for each selected frame and creates feature vectors based on the distances between pairs of landmarks. Each feature vector is normalized using the reference feature vector, which is generated from the distances in the reference frame. These vectors are then stacked into a matrix and used as input to a Convolutional Neural Network (CNN). Camera Close-Up [14] adapts this work by adopting a frame selection based on bins and altering the CNN design while maintaining the matrix of feature vectors based on landmark distances. These minor changes are sufficient to outperform the earlier method using the data collected for their study. To the best of our knowledge, these are the most recently published methods addressing the close-up paradigm for liveness classification.

III. METHODOLOGY

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 [14], sharing similarities in the frame selection process and some portions of its architecture. The Hybrid Close-Up approach integrates the concept of projective invariants, as defined by De Riccio et al. [46], 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 II-C provides a general overview of the method, and each module is described in this section.

A. Frame selection

The first module of the pipeline is responsible for selecting the frames used in the next steps and is identical to the Camera Close-up frame selection process. First, a face detector and a landmark extractor are used to discard frames that do not contain a face or where landmarks cannot be computed. Next, N frames are selected using a system of s bins, where each frame is assigned to a bin based on its timestamp, starting with the face farthest from the camera and ending with the face closest to the camera. Then, the frame closest to the center of the video is sampled to be a reference frame, and $\frac{N}{s}$ frames are randomly chosen from each bin.

B. Feature extraction

The next module is responsible for extracting features from the selected frames.

The Hybrid Close-Up model has two types of features: distortion features and frame embeddings. Distortion features are computed by first extracting landmarks from the faces in the frames and then calculating the Euclidean distances between all pairs of landmarks, excluding the distances between landmarks in the mouth region. This process produces for each selected frame-k a distortion feature vector $d_k = (f_{k0}, f_{k1}, \ldots, f_{kM-1})$ with $k \in [0..N-1]$ of length M, where M is the number of distances between pairs of landmarks.

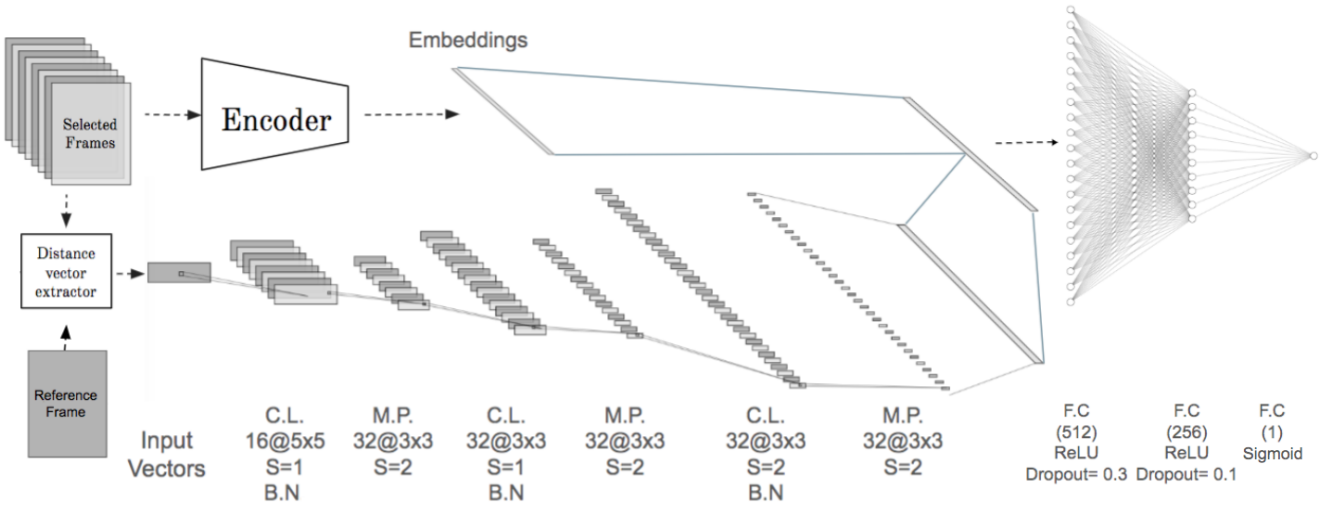


Fig. 2. Hybrid CloseUp scheme; C.L:2D Convolutional Layer; M.P: Max-Pooling Layer; F.C: Fully Connected; S: stride; B.N: Batch Normalization

Similarly, a reference distortion feature vector $d_{ref} = (f_{ref0}, f_{ref1}, \ldots, f_{refM-1})$ is computed using the reference frame. For each d_k, where $k \in [0..N-1]$, a normalized distortion feature vector d_{nk} is calculated as

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

Lastly, the distortion feature vectors are reorganized to form an $N \times M$ distortion feature matrix.

The second type of feature is frame embeddings. These are extracted using the encoder of a ResNet model pre-trained on ImageNet, which computes the embeddings of lenght E for each selected frame, except for the reference frame. The embeddings are concatenated producing an video embedding vector of lenght $N \times E$.

By following these steps, the module effectively extracts and normalizes the required features from the frames.

C. Classification

The final module of the Hybrid Close-Up method is responsible for classifying between live and spoof using the computed features. The parameters of the proposed CNN are shown in Figure II-C. The distortion feature matrix serves as the input to the convolutional layers of the Hybrid Close-up model, which consists of only three convolutional layers, each followed by batch normalization, ReLU activation function, and max pooling.

The video embedding vector is concatenated with the flattened features produced by the convolutional layers and the resulting linear feature vector serves as input to a fully connected network with two hidden layers, followed by the final layer consisting of a single neuron with a sigmoid activation function.

IV. Experiments and Results

This section is dedicated to describing the setup used to conduct the experiments and their results. We implemented the Camera Close-up and Face Close-up methods based on the descriptions available in their respective articles using PyTorch and these implementations are available at https://github.com/BOVIFOCR/Active_liveness-Close_Up_methods .

A. Experiment Setting

The number of video bins s used in the frame selection module of the Hybrid Close-up method was set to 3. The Python module of dlib library [47] was utilized for face detection and landmark extraction in all implemented models.

The used landmark extractor computes 68 points, with 10 of them being from the mouth region. Thus, the distortion feature vector 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). For embedding extraction in the Hybrid CLose-up method, the encoders from ResNet-18 and ResNet-50 were used, producing for each selected frame embeddings of sizes 512 and 2048, respectively. Thus, the final frame embedding vector that is concatenated with the features from convolutional layers has dimension $N \times 512$ when using the ResNet-18 encoder and $N \times 2048$ with the ResNet-50 encoder. It is important to note that these encoders were pre-trained on ImageNet, and their weights were frozen during the training of the remaining Hybrid Close-up architecture on our dataset.

We used the number of selected frames for the Camera Close-up and Face Close-up methods as indicated in their respective papers from the best results achieved there. Initially, these active models were compared to ResNet-18 and ResNet-50, which used a single frame randomly sampled from the input video and with encoders pre-trained on ImageNet. Since they rely on only one frame for liveness classification and do not leverage the close-up movement of the dataset, the ResNet models in this work operate under conditions similar to passive approaches based on single-frame classification. Thus, they represent a naive passive approach. Then, we also employ a

majority vote scheme on the ResNet models aiming for a fair comparison with the active ones.

The dataset currently used in the experiments contains 683 live samples and 1027 spoof samples, which are divided into training, validation, and test partitions in proportions of 60%, 20%, and 20%, respectively. The results present the evaluation of Accuracy, HTER (Half Total Error Rate), and F1-score on the test set, using the trained weights that achieved the lowest HTER on the validation set. All experiments were run 10 times, and the reported values are the averages followed by the standard deviations of these runs. Every network was trained using the ADAM optimizer, with a learning rate of 0.001, a batch size of 50, and 500 epochs.

Firstly, experiments were conducted to evaluate the impact of the number of selected frames N in the proposed method as shown in Table II. It can be seen that the Hybrid Close-up model with the ResNet-18 encoder improves its results by increasing the number of selected frames whilst it reaches its maximum efficiency at 18 frames when using the ResNet-50 encoder. In light of the presented results, the following experiments were conducted selecting 18 and 30 frames when applying ResNet-50 and ResNet-18, respectively.

TABLE II
PERFORMANCE IMPACT COMPARISON

Encoder	N	Accuracy(%)	HTER(%)	F1-score
ResNet-18	12	97.04 ± 0.72	3.23 ± 0.81	0.962 ± 0.009
	18	96.75 ± 0.86	3.63 ± 0.99	0.958 ± 0.011
	24	96.75 ± 0.55	3.65 ± 0.63	0.958 ± 0.006
	30	97.16 ± 0.65	3.08 ± 0.92	0.964 ± 0.008
ResNet-50	12	97.99 ± 0.57	2.27 ± 0.79	0.976 ± 0.008
	18	98.11 ± 0.68	2.12 ± 0.79	0.976 ± 0.008
	24	97.87 ± 0.13	2.39 ± 0.11	0.973 ± 0.001
	30	97.93 ± 0.47	2.37 ± 0.54	0.973 ± 0.006

Additionally, Table III shows a comparison of the studied methods. We start with the performance of the active baselines Face Close-Up and Camera Close-Up. And, we also display the results achieved by ResNet-50 and ResNet-18 using both only a single image and a majority voting scheme from the individual classification of 18/30 randomly sampled frames. Finally, we show the results for the proposed Hybrid Close-Up using the ResNet-50 encoder to extract frame embeddings.

TABLE III
STATE OF THE ART PERFORMANCE COMPARISON

Method	Accuracy(%)	HTER(%)	F1-score
Face Close-up [13]	85.03 ± 1.06	15.22 ± 1.20	0.817 ± 0.014
Camera Close-up [14]	91.36 ± 0.49	8.66 ± 0.72	0.890 ± 0.007
ResNet-18 (single-frame)	94.44 ± 0.44	5.85 ± 0.55	0.930 ± 0.005
ResNet-18 (majority voting - 30)	96.48 ± 0.62	3.08 ± 0.92	0.964 ± 0.008
ResNet-50 (single-frame)	95.78 ± 0.23	4.16 ± 0.17	0.952 ± 0.002
ResNet-50 (majority voting - 18)	98.81 ± 0.46	1.57 ± 0.31	0.982 ± 0.004
Hybrid Close-up	98.11 ± 0.68	2.12 ± 0.79	0.976 ± 0.008

We observe that the ResNet models using a single frame outperform the active baselines in all three reported metrics. Moreover, the Hybrid Close-up method by combining the distortion features of Camera Close-up and embeddings from the ResNet-50 encoder improves the previous results by a significant margin. However, the best values were achieved using a majority voting scheme of the ResNets results, highlighting the current gap between the latest active methods and passive approaches.

We hypothesized that the large size of the distortion feature (50,144/36,416 for Camera Close Up/Face Close Up when compared to the one of the ResNet 50/18, i.e., 2048/512) strongly impacted the classification performance.

V. CONCLUSION

In this work, we proposed a method for active face anti-spoofing based on the close-up face challenge paradigm. The method leverages the fusion of features extracted from face embeddings and distances between landmarks, thus exploiting temporal and spatial features enhanced by the dynamic aspect introduced by the active interaction with the user at the moment of sample capture. Furthermore, Hybrid Close-Up outperforms the latest and most relevant methods proposed in the area that use the same paradigm in a diverse and unconstrained dataset. Nevertheless, the existing disparity between active liveness detectors and their passive counterparts is still latent as demonstrated in this paper.

Future works are encouraged to explore more robust frame selection criteria by combining the temporal information with the spatial content on each image, leading to a selection of frames with more relevant information.

Another promising study from our hypothesis presented in the experiments is to develop a smaller and more suitable distortion embedding to be fused with the ones from ResNet-18 and ResNet-50. We also plan to employ an optical flow approach for the close-up challenge as already employed on other works [48], [49].

ACKNOWLEDGMENT

This work was supported by a tripartite-contract, i.e., unico - idTech, UFPR (Federal University of Paraná), and FUNPAR (Fundação da Universidade Federal do Paraná). We also thank the National Council for Scientific and Technological Development (CNPq) (# 315409/2023-1) for supporting Prof. David Menotti.

REFERENCES

[1] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5609–5631, 2022.

[2] D. Gollmann, "Computer security," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 544–554, 2010.

[3] S. Z. S. Idrus, E. Cherrier, C. Rosenberger, and J.-J. Schwartzmann, "A review on authentication methods," *Australian Journal of Basic and Applied Sciences*, vol. 7, no. 5, pp. 95–107, 2013.

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

[5] M. Killioğlu, M. Taşkiran, and N. Kahraman, "Anti-spoofing in face recognition with liveness detection using pupil tracking," in *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2017, pp. 000 087–000 092.

[6] M. Singh and A. Arora, "A robust anti-spoofing technique for face liveness detection with morphological operations," *Optik*, vol. 139, pp. 347–354, 2017.

[7] M. Mohzary, K. J. Almalki, B.-Y. Choi, and S. Song, "Your eyes show what your eyes see (y-eyes): Challenge-response anti-spoofing method for mobile security using corneal specular reflections," in *1st Workshop on Security and Privacy for Mobile AI*, 2021, p. 25–30.

[8] Y. Liu, Y. Tai, J. Li, S. Ding, C. Wang, F. Huang, D. Li, W. Qi, and R. Ji, "Aurora guard: Real-time face anti-spoofing via light reflection," *CoRR*, vol. abs/1902.10311, 2019. [Online]. Available: http://arxiv.org/abs/1902.10311

[9] J. M. D. Martino, Q. Qiu, T. Nagenalli, and G. Sapiro, "Liveness detection using implicit 3D features," *CoRR*, vol. abs/1804.06702, 2018. [Online]. Available: http://arxiv.org/abs/1804.06702

[10] W. Xu, J. Liu, S. Zhang, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "Rface: Anti-spoofing facial authentication using cots rfid," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021, pp. 1–10.

[11] M. Ezz, M. Ayman, and A. Elshenawy Elsefy, "Challenge-response emotion authentication algorithm using modified horizontal deep learning," *Intelligent Automation and Soft Computing*, vol. 35, pp. 3659–3675, 09 2022.

[12] Z. Ming, J. Chazalon, M. M. Luqman, M. Visani, and J.-C. Burie, "Facelivenet: End-to-end networks combining face verification with interactive facial expression-based liveness detection," in *2018 24th Int. Conf. on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3507–3512.

[13] Y. Li, Z. Wang, Y. Li, R. Deng, B. Chen, W. Meng, and H. Li, "A closer look tells more: A facial distortion based liveness detection for face authentication," in *Asia CCS '19: ACM Asia Conference on Computer and Communications Security*, 07 2019, pp. 241–246.

[14] A. Castelblanco, E. Rivera, J. Solano, L. Tengana, C. López, and M. Ochoa, "Dynamic face authentication systems: Deep learning verification for camera close-up and head rotation paradigms," *Comput. Secur.*, vol. 115, no. C, apr 2022.

[15] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 504–517.

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

[17] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR Int. Conf. on Biometrics (ICB)*, 2012, pp. 26–31.

[18] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG - Int. Conf. of Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 1–7.

[19] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[20] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2268–2283, 2016.

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

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

[23] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4675–4684.

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

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

[27] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "CelebA-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 70–85.

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

[29] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 141–145, 2017.

[30] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Lbp-top based countermeasure against face spoofing attacks," in *Computer Vision - ACCV 2012 Workshops*, J.-I. Park and J. Kim, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 121–132.

[31] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

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

[33] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, "Dual-cross central difference network for face anti-spoofing," *CoRR*, vol. abs/2105.01290, 2021. [Online]. Available: https://arxiv.org/abs/2105.01290

[34] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," *CoRR*, vol. abs/2003.04092, 2020.

[35] Q. Zhou, K.-Y. Zhang, T. Yao, X. Lu, R. Yi, S. Ding, and L. Ma, "Instance-aware domain generalization for face anti-spoofing," 2023.

[36] B. M. Le and S. S. Woo, "Gradient alignment for cross-domain face anti-spoofing," 2024.

[37] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.

[38] J. Wang, J. Zhang, Y. Bian, Y. Cai, C. Wang, and S. Pu, "Self-domain adaptation for face anti-spoofing," *CoRR*, vol. abs/2102.12129, 2021. [Online]. Available: https://arxiv.org/abs/2102.12129

[39] M. Singh and A. S. Arora, "A novel face liveness detection algorithm with multiple liveness indicators," *Wireless Personal Communications*, vol. 100, no. 4, pp. 1677–1687, Jun 2018.

[40] I. Sluganovic, M. Roeschlin, K. Rasmussen, and I. Martinovic, "Using reflexive eye movements for fast challenge-response authentication," in *CCS '16: ACM SIGSAC Conference on Computer and Communications Security*, 10 2016, pp. 1056–1067.

[41] M. Shen, Z. Liao, L. Zhu, R. Mijumbi, X. Du, and J. Hu, "Iritrack: Liveness detection using irises tracking for preventing face spoofing attacks," 2018. [Online]. Available: https://arxiv.org/abs/1810.03323

[42] P. McShane and D. Stewart, "Challenge based visual speech recognition using deep learning," in *2017 12th Int. Conf. for Internet Technology and Secured Transactions (ICITST)*, 2017, pp. 405–410.

[43] E. Uzun, S. Chung, I. Essa, and W. Lee, "rtCaptcha: A real-time captcha based liveness detection system," in *Conference: The Network and Distributed System Security Symposium (NDSS)*, 02 2018.

[44] C.-L. Chou, "Presentation attack detection based on score level fusion and challenge-response technique," *The Journal of Supercomputing*, vol. 77, 05 2021.

[45] M. De Marsico, M. Nappi, D. Riccio, and J.-L. Dugelay, "Moving face spoofing detection via 3D projective invariants," in *2012 5th IAPR Int. Conf. on Biometrics (ICB)*. IEEE, 2012, pp. 73–78.

[46] D. Riccio and J.-L. Dugelay, "Geometric invariants for 2d/3d face recognition," *Pattern Recognit. Lett.*, vol. 28, pp. 1907–1914, 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:5490223

[47] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[48] L. Li, Z. Xia, J. Wu, L. Yang, and H. Han, "Face presentation attack detection based on optical flow and texture analysis," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1455–1467, 2022.

[49] S. Chen, A. Pande, and P. Mohapatra, "Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones," in *12th Annual International Conference on Mobile Systems, Applications, and Services*, 2014, pp. 109–122.