Temporal Similarity Analysis of Remote Photoplethysmography for Fast 3D Mask Face Presentation Attack Detection

Si-Qi Liu Xiangyuan Lan Pong C. Yuen
Department of Computer Science, Hong Kong Baptist University, Hong Kong
siqiliu, lanxiangyuan, pcyuen@comp.hkbu.edu.hk

Abstract

To tackle the 3D mask face presentation attack, remote Photoplethysmography (rPPG), a biomedical technique that can detect heartbeat signal remotely, is employed as an intrinsic liveness cue. Although existing rPPG-based methods exhibit encouraging results, they require long observation time (10-12 seconds) to identify the heartbeat information, which limits their employment in real applications such as smartphone unlock and e-payment. To shorten the observation time (within 1-second) while keeping the performance, we propose a fast rPPG-based 3D mask presentation attack detection (PAD) method by analyzing the similarity of local facial rPPG signals in the time domain. In particular, a set of temporal similarity features of facial and background local rPPG signals are designed and fused to adapt the real world variations based on rPPG shape and phase properties. For better evaluation under practical variations, we build the HKBU-MARsV2+ dataset that includes 16 masks from 2 types and 6 lighting conditions. Finally, extensive experiments are conducted on 11092 short-term video slots from 4 datasets with a large number of real-world variations, in terms of mask type, lighting condition, camera, resolution of face region, and compression setting. Results show that the proposed TSrPPG outperforms the state-of-the-art competitors dramatically on discriminability and generalizability. To our best knowledge, this is the first work that addresses the length of observation time issue of rPPG-based 3D mask PAD.

1. Introduction

Face recognition technique has been employed in various applications nowadays due to its practicability and convenience, especially the access control and E-payment. Naturally, the security issue of face recognition systems becomes a critical concern. Face recognition systems are vulnerable to presentation attacks when obtaining one's face image becomes easier with the increasing popularity of social networks. Face presentation attacks can be conducted at very low cost with prints or screen. To detect them, great efforts have been made in the last decades [25, 22, 7, 34, 32, 5, 11] and a number of liveness cues have been proposed and studied, such as the appearance-based cues including texture [22, 5], image quality [11], reflection patterns [32, 21], and motion-based cues including eyes or mouth motion [25] and facial expression [7].

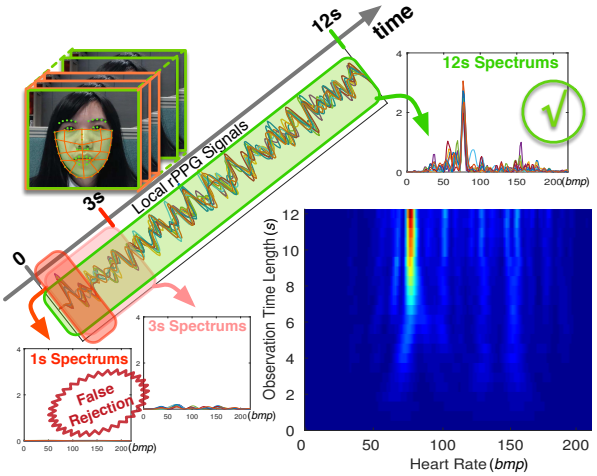


Figure 1. An example of rPPG frequency analysis when the observation time becomes short. Local rPPG signals are extracted from local face regions that are located with facial landmarks. With longer observation time (12 seconds in the green window), the signal spectrum shows a clear peak at the heartbeat frequency. With limited observation time (3 seconds and 1 second in red windows), the heartbeat sign becomes less distinguishable which causes false rejection error. The right subfigure visualizes the spectrum variation when the observation time becomes shorter. Each row in the matrix represents the average of multiple local rPPG spectrums given a specific observation time length.

In addition to the two conventional attacks, 3D mask attack has attracted increasing attention since the customized 3D mask can be easily made at an affordable price [10]. Although the appearance-based methods exhibit strong ability

on detecting Thatsmyface masks[1] with 3D printing quality defect [10], their effectiveness can not stand when encountering high quality 3D masks with vivid texture and shape as real faces [18]. In addition, the appearance-based methods fail when the training data environment (light, camera, and spoofing media) is different from the one of testing data due to their data-driven nature [32, 18].

Recently, a new liveness cue based on remote photoplethysmography (rPPG) is proposed to tackle the 3D mask attack challenge [18, 16]. rPPG is a new technique that can extract the heartbeat signal through normal RGB camera by measuring the subtle skin color variations caused by the blood pulse. For 3D mask attack problem, such a liveness signal can only be observed on genuine faces but not on masked faces because the 3D mask blocks the light transmission from the facial skin [18]. Since the blocking of heartbeat signal is independent to the mask appearance, the rPPG-based 3D mask PAD methods can detect the hyper real mask well and shows good generalizability [18, 19].

However, such good performance is constrained on long observation time (around 10-12 seconds on existing datasets), which is not fast enough for practical application usage such as the smart-phone access or the electronic payment. The long-term waiting will make costumer anxious and destroy the user experience of the entire system. For existing rPPG-based methods that rely on signal spectrum analysis in the frequency domain, the input face video should be long enough to contain sufficient number of stable heartbeat cycles to obtain distinguishable response (peak) at the heartbeat frequency [18, 16]. Also, since rPPG measurement relies on the subtle skin color variation which is sensitive to lighting condition and facial movement [8], it requires even longer observation time in unconstrained application scenarios. Otherwise, false rejection error increases since the rPPG spectrum on genuine faces and masked faces can hardly be differentiated. Figure 1 shows an example of the rPPG spectrum analysis with different length of observation time under room light.

To shorten the observation time while achieving better the performance, we propose the **t**emporal **s**imilarity of **rPPG** (TSrPPG) method for fast 3D mask PAD. Different from the spectrum analysis approach which exploits the heartbeat from long-term observation in frequency domain [18, 16], the proposed TSrPPG can extract the heartbeat vestige within 1 second, by analyzing the rPPG signal waveform in time domain. Within the proposed TSrPPG, the similarities between the local rPPG signals extracted from facial and background regions are employed as the liveness discriminant, based on the property that facial rPPG signals are all generated from the same source — the heartbeat. In particular, an rPPG temporal similarity feature operator is proposed to extract both the similarity and the amplitude information in short observation time. On top of that, a set of distance metric is designed according to the waveform characteristic of local rPPG signals, in terms of their shapes and phases. The final result is obtained through score-level-fusion to better adapt different variations in practice. The proposed TSrPPG can be regarded as a general framework that allows different types of signal similarity metric and fusion strategies.

Moreover, to better evaluate the rPPG-based approach with practical variations, this paper extends the HKBU-MARsV2 dataset [17] into a larger scale by adding 4 subjects with their customized hyper real REAL-f masks and increasing the number of videos by 3 times. The new dataset HKBU-MARsV2+ now contains 16 subjects with 10 hyper real REAL-f masks and 6 Thatsmyface masks, 6 lighting conditions, and 480 videos in total.

In summary, the contributions of this paper are:1) A fast rPPG-based 3D mask PAD method based on time domain rPPG signal analysis. 2) An rPPG feature operator with a set of time domain rPPG similarity measurement, in terms of their signal shapes and phases. 3) An new dataset which contains 2 types of masks, 6 lighting conditions, and 16 subjects. Extensive experiments on 4 datasets with a large number of variations (mask type, light, and video quality) demonstrate that the proposed method improves the performance dramatically with short observation time compared with the state-of-the-art rPPG-based methods.

To our best knowledge, this is the first work that addresses the long observation time issue in the typical rPPG-based 3D mask PAD methods.

2. Related Work

Existing face presentation attack detection (PAD) methods can be mainly classified into appearance-based approach, motion-based approach and rPPG-based approach based on the liveness cues.according to the liveness cues.

Appearance-based Approach. Due to the precision of 3D printing, multi-scale LBP [22, 5] and other texture features [1] show the ability to identify the detailed texture differences between mask and real face. However, it is also found that they fail on hyper real 3D masks [18, 19] since the quality defects of texture and color can be imperceptible. The image distortion analysis [11] based on the quality defects of the spoofing instrument [13], *e.g.*, color diversity, the reflection patterns [32, 21], or the Moiré patterns [26] also facing this challenge since mask may not contain those quality defects. Besides, they expose limited generalizability when camera or light settings varies [32, 17]. When deep learning is getting popular recently, deep features have also been employed in face PAD and exhibit outstanding discriminability [23, 33]. Using auxiliary information such as the facial depth map [20] or the 3D structure [12] to supervise the learning of deep network can further boost the per-

[1] www.thatsmyface.com

formance. Still, the over-fitting problem remains unsolved due to the intrinsic data-driven nature.

Near-infrared or thermal camera can be effective as the plastic material blocks the heat radiation [3]. However, it requires additional devices which is not suitable for existing well-established RGB camera based systems.

Motion-based Approach. The motion-based approach is based on motion difference between genuine faces and static spoofing materials. As such, eye-blink [25] or mouth movement [7] described using the optical flow can be used to differentiate 2D spoofing media from 3D genuine faces. Facing 3D mask attacks, these methods may not work since the aforementioned motion pattern can be well preserved on 3D masks with exposed eyes and mouth [9]. In addition, the soft silicon gel mask that preserves the subtle facial muscle motion makes the motion cue more unreliable.

rPPG-based Approach. rPPG is a new biomedical technique that can measure human heartbeat remotely through a normal RGB camera based on the principle of contact-PPG [27, 29, 15]. When applying rPPG on face PAD, 3D masks on live faces blocks the heartbeat signals so that attacks can be detected by analyzing the spectrum of observed signals strength [16]. Local rPPG extracted from local facial regions can provide more spatial information which is more robust to cross-dataset testing [18]. On top of that, they developed a new version based on correlation filter which is robust to noisy spectrums [19]. In addition to 3D mask attacks, the rPPG-based solution can be effective on conventional prints and screen attacks [24] because these materials block the heartbeat signals in the same way [16]. The rPPG-based 3D mask PAD methods exhibit encouraging performance on 3DMAD [10] and HKBU-MARsV1+ [19]. However, the reported results are based on long-term input face videos (10s and 12s for 3DMAD and HKBU-MARsV1+), which is not applicable in practice.

3. Analysis of rPPG-based 3D mask PAD

Photoplethysmography (PPG) is a heartbeat monitoring technique which uses pulse oximeter to illuminate the skin and measure the changes in light absorption caused by the pumping of blood during cardiac cycles [30]. Remote PPG (rPPG) follows the same principle while using normal RGB camera to measure the heartbeat-caused skin color variation under room lighting condition. As a liveness cue for face anti-spoofing, for a genuine face, the heartbeat cycles can be detected from the observed rPPG signals. For a masked face, the 3D mask material blocks the light transmission and rPPG signals only remain environmental noise [18].

Since the heartbeat is periodic, it is intuitive to analyze the observed rPPG signal in the frequency domain. The liveness evidence can be measured by the signal strength such as the maximum amplitude, or the signal to noise ratio (SNR) of the rPPG spectrum. Following this approach, the rPPG-based solution has shown promising results under both intra and cross dataset scenarios [18, 16]. However, existing methods do not consider the observation time issue of a face recognition system and their reported results are restricted to 10-12 seconds videos, which is not fast enough for real application.

To obtain an rPPG spectrum with clear heartbeat periodicity, we empirically found that the input face video should contain at least 3-5 heartbeat cycles, under well-controlled lighting condition. Ideally, the observation time is 3-5 seconds for a static user whose heartbeat is around 60 beats per minute (bpm). While under unconstrained environment in real applications, the heartbeat sign in the rPPG spectrum can easily get contaminated since the rPPG signal, *i.e.*, the subtle skin color variation is sensitive to camera settings and lighting condition based on the principle discussed above. Therefore, for existing rPPG-based PAD method, longer input video length is obligatory to obtain more stable rPPG signals with distinguishable periodicity. Otherwise, false rejection error occurs as the rPPG spectrum on genuine faces and masked face can hardly be differentiated. This will make the entire system less convenient especially for the application that requires quick response time, *e.g.*, the mobile phone unlock or E-payment. Fig. 1 shows an example of this deduction. Given a genuine face video recorded through a web-cam under room light[2], the periodicity of the rPPG spectrums tends to be less significant when the observation time (rPPG signal length) becomes shorter. Note that the heartbeat peak is obvious on 12 seconds rPPG signals but concealed on 3 seconds and 1 second signals. The trend of spectrum variation with different observation time length also indicates the performance degradation of rPPG frequency analysis based 3D mask PAD methods when observation time becomes limited.

4. Proposed Method

To overcome the limitations of long observation time, we propose a fast rPPG-based 3D mask PAD solution by analyzing the temporal similarity of local rPPG signals. For preprocessing, we first use the CLNF landmark tracker [2] to obtain 68 points facial landmarks and define local facial regions $r_1, r_2, \ldots, r_N$ on them. Then, local rPPG signals $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N$ are extracted through CHROME [8], where $\boldsymbol{s}_i = [s_i^1, \ldots, s_i^T]$ and T is the number of frames. Implementations details are summarized in Section 5.

4.1. Temporal rPPG Similarity Feature for Fast 3D Mask PAD

Given local rPPG signals $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N]$, it is intuitive to use the amplitude or the signals themselves as features to detect masks from genuine face since one contains

[2]Example face video in Fig. 1 is selected from HKBU-MARsV1 dataset [18]

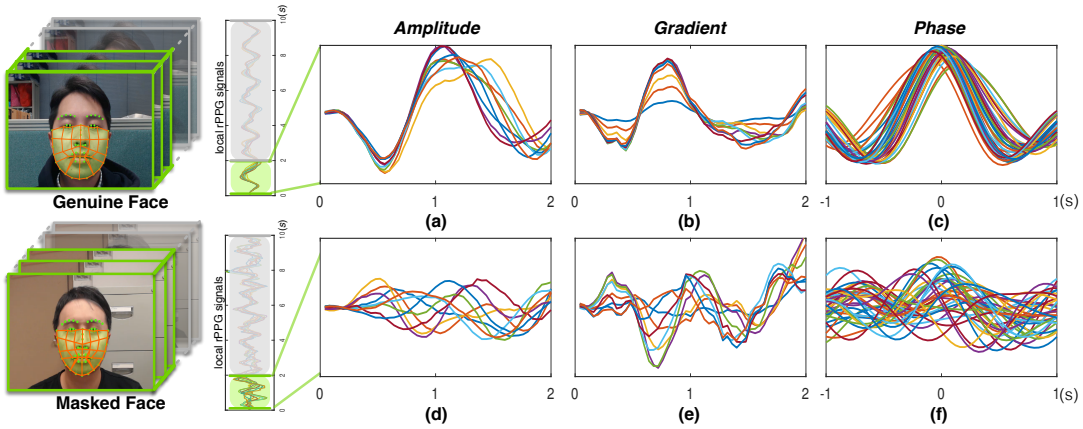


Figure 2. Similarity comparison of local rPPG signals between genuine face and masked face, in terms of the amplitude, gradient and phase.

the heartbeat information and the other is random noise. However, this approach may not work because of the following issues. i) The rPPG signal amplitude for different subjects varies due to the difference of heartbeat strength or skin color. ii) Even for subjects with similar heartbeat strength, the phase of their rPPG signals will vary from time to time. iii) Moreover, in unconstrained environment with lower light or poor quality camera, rPPG signals on masked faces and genuine faces may exhibit similar amplitude and result in false acceptance or false rejection errors.

Therefore, to avoid the variation of rPPG signals amplitude and phase for different subject, we tackle the problem by analyzing the temporal similarity between local rPPG signals for each subject. Based on the principle that rPPG signals are generated from the human heartbeat, given a genuine face, the local rPPG signals extracted from different facial regions should be similar, in terms of their shape, phase and amplitude. While for a masked face, the observed local rPPG signals vary randomly since their sources are from the environmental noise. This indicates that the similarity between local rPPG signals of each subject can be used as the liveness cue for differentiating 3D mask attacks from genuine faces.

TSrPPG Feature Operator One intuitive solution to measure the similarity is using the distance metric $d(\boldsymbol{s}_i, \boldsymbol{s}_j)$. However, $d(\boldsymbol{s}_i, \boldsymbol{s}_j)$ from genuine faces and masked faces can be similar since the amplitude of rPPG signals on masked faces is smaller than those on genuine faces [18] (Defect of this solution is illustrated in experiment by the TSrPPG-strfwd method). In stead of directly measure the distance, we propose the temporal rPPG feature operator as follows:

$$\boldsymbol{TSrPPG_{i,j}}[m] = \int_{-\infty}^{+\infty} \mathcal{D}(\boldsymbol{s}_i[t], \boldsymbol{s}_j[t+m])dt \quad (1)$$

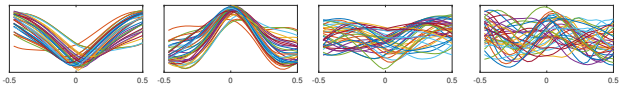
where t is the observation time, m is the shifting index, and $\mathcal{D}(\boldsymbol{s}_i, \boldsymbol{s}_j)$ measures the similarity of $\boldsymbol{s}_i, \boldsymbol{s}_j$. $\boldsymbol{TSrPPG}$ is first calculated for all $\boldsymbol{TSrPPG_{i,j}}$ where $i <= j, i, j = 1, \ldots, N$. Then, the mean, standard deviation of columns, and the center value ($n = 0$) of rows is concatenated as the TSrPPG feature.

The example of $\boldsymbol{TSrPPG}$ on genuine and masked face is visualized in Fig. 3(a) and Fig. 3(c). For genuine face, the center value $TSrPPG_{i,j}[0]$ measures the distance $\mathcal{D}$ of $\boldsymbol{s}_i, \boldsymbol{s}_j$ when the two signals are aligned. When m comes to half of the heartbeat cycle $H/2$, the largest distance which reflects the maximum amplitude information is acquired. When $|m|$ is between 0 to $H/2$, the detailed heartbeat information hidden in the distance variation of $\boldsymbol{s}_i$ and shifted $\boldsymbol{s}_j$ is obtained. Therefore, liveness pattern can be extracted between half to one heartbeat cycle (around 1 second) using the proposed TSrPPG feature operator. While for masked face, such pattern differs from the one on genuine face as the $\boldsymbol{s}_i, \boldsymbol{s}_j$ may not align at $m = 0$ and maximum amplitude may not be reached at $|m| = H/2$.

rPPG from Background Region As indicated above, for masked faces, the facial rPPG signals should be identical to the rPPG signals extracted from background since both of them are random noise. While for genuine faces, they should be less similar since facial rPPG signals contains heartbeat information. Inspired by this property, we further boost the discriminability by extracting the TSrPPG feature between local facial rPPG signals and background signals as follows:

$$\boldsymbol{TSrPPG}^{bg}_{\boldsymbol{i,j}}[m] = \int_{-\infty}^{+\infty} \mathcal{D}(\boldsymbol{s}^f_i[t], \boldsymbol{s}^b_j[t+m])dt \quad (2)$$

where $\boldsymbol{s}^f_i, \boldsymbol{s}^b_j$ is the rPPG signals extracted from local facial regions and background regions respectively. The $\boldsymbol{TSrPPG}^{bg}$ is also obtained by getting all $\boldsymbol{TSrPPG}^{bg}_{\boldsymbol{i,j}}$, where $i = 1, \ldots, N$, and $j = 1, \ldots, J$. The background TSrPPG feature is then constructed as the concatenation of the mean and standard deviation of each column, and the center value ($n = 0$) of each row.



(a) Amplitude (b) Gradient (c) Amplitude (d) Gradient

Figure 3. Example of TSrPPG matrix on genuine face (left two) and masked face (right two)

We construct the final liveness feature as the concatenation of TSrPPG feature from local facial regions and TSrPPG feature between local facial regions and background regions.

To thoroughly extract the liveness information from local rPPG signals, we design three types of similarity measurement based on their shapes and phases. Fig. 2 shows an example of the temporal similarities of local rPPG signals of genuine face and masked face.

Amplitude. The similarity of the amplitude can reflect the liveness evidence because the amplitude of rPPG signals on genuine faces varies along local facial regions and forms a stable distribution for different subjects. This is due to the following two aspects:

1) The density of facial blood vessels forms a stable spatial distribution for different subjects [6]. Since the rPPG signal is calculated from the color variation of blood, the signal strength highly depends on the density of blood vessels. For different genuine faces, the cheek and chin with dense arterial and venous blood vessels [6] always yield significant rPPG amplitude [14]. 2) The transmittance of facial skin also forms a stable pattern for different frontal faces. Based on the physical model of rPPG [8], larger light incident angle brings weaker pulse signal observation so the transmittance varies along facial structures according to its angle towards the camera. The bridge of the nose and the face border has a larger angle than the cheek and chin so the amplitude of rPPG signals within those regions is lower. As a result, the similarity of the amplitude of local rPPG signals also contains such property and the final similarity feature encodes this spatial information. Such spatial distribution does not hold on masked faces where their source is environmental noise.

Given local rPPG signals $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N]$, we measure the similarity of each two signal $\boldsymbol{s}_i, \boldsymbol{s}_j$ with Euclidean distance

$$\mathcal{D}_a(\boldsymbol{s}_i, \boldsymbol{s}_j) = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2 \tag{3}$$

The amplitude similarity feature can be robust to different lighting conditions since the amplitude of local rPPG signals are globally affected by such variation and thereby maintain a stable similarity pattern.

Gradient. Inspired by the use of edge or shape in designing 2D image features (*e.g.*, SIFT, HOG), the gradient of rPPG signals implies the detailed shape information of the heartbeat such as the slope of the tangent (shown in Fig. 2 b). Hence, we take the gradient of an rPPG signal which contain finer information of the heartbeat waveform. In practice, we use the Gaussian Derivative Filter (GDF) to suppress the high frequency noises and artifacts while preserving the detailed pulse waveform. The Gaussian kernel $G[n]$ with size L is constructed as

$$G[l] = e^{-\frac{(l-(L/2))^2}{2\sigma^2}} \tag{4}$$

where σ is the standard deviation. Then, the Gaussian derivative kernel is computed as $F[l] = G[l+1] - G[l]$ where $l = 1, 2, 3, \ldots, L-1$. The Gaussian derivative of rPPG signal $\boldsymbol{s}_i^f[t]$ is calculated by convolving the original rPPG signal $\boldsymbol{s}_i[t]$ with the kernel $F[t]$.

Since the gradient of rPPG signal tends to amplify the finer shape, it may also be sensitive to interferences in unconstrained scenarios. For better robustness we use the normalized cross correlation (NCC) metric to measure the shape similarity of $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$

$$\mathcal{D}_s(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{(\boldsymbol{s}_i^f(t) - \bar{\boldsymbol{s}}_i^f)^\intercal(\boldsymbol{s}_j^f(t) - \bar{\boldsymbol{s}}_j^f)}{\|\boldsymbol{s}_i^f(t) - \bar{\boldsymbol{s}}_i^f\|_2 \|\boldsymbol{s}_j^f(t) - \bar{\boldsymbol{s}}_j^f\|_2} \tag{5}$$

where $\bar{\boldsymbol{s}}^f$ is the mean of filtered rPPG signal $\boldsymbol{s}^f$. Example of TSrPPG pattern with $\mathcal{D}_s$ is visualized in Fig. 3(b) and Fig. 3(d) for genuine and masked face.

Phase. Despite the shape information described through amplitude and gradient, the phase (time) delay of local rPPG signals also encloses liveness information. For rPPG signals on genuine faces, they share the close phases (Fig. 2 (a)) since the blood flow through different facial regions at the same time. While for rPPG signals on masked faces, their phases are inconsistent since random noises are from various sources. Fig. 2 (c) and 2(f) show an example of their difference. We measure phase similarity of two rPPG signals as the lag argument of the maximum response of their cross-correlation. So the distance metric is defined as,

$$\mathcal{D}_p(\boldsymbol{s}_i, \boldsymbol{s}_j) = \boldsymbol{s}_i \cdot \boldsymbol{s}_j \tag{6}$$

4.2. Classification

Given the 3 sets of local rPPG similarity features, we use SVM with RBF kernel to obtain their classification scores s_a, s_s and s_p. Since each similarity feature is based on different properties of rPPG signals, they are effective to handle different variations in real world environment. Therefore, score level fusion is employed to summarize them so that each can contribute appropriately to the final decision. In this paper, the classification score s_a, s_s and s_p are weighted (the classification results (AUC) in training stage) summed as the final score.

5. Experiments

5.1. Experimental Setup

We evaluate our method on 3DMAD [9] dataset that contains 17 subjects with customized Thatsmyface (TMF) [3]

[3] http://thatsmyface.com/

	Subject/Mask Num.	Video Slot Num.	3D Mask Type	Light Cond.	Camera	Face Resolution (pixel)	Compression
3DMAD [18]	17 17	2550	TMF	1(Studio)	Kinect	80×80	Motion JPEG
HKBU-MARsV1+ [19]	12 12	2160	TMF+RF	1(Room)	Logitech C920	200×200	H.264
CSMAD [3]	14 6	1582	Silicon	4	RealSense SR300	350×350	H.264
HKBU-MARsV2+	16 16	4800	TMF+RF	6	MV-U3B	200×200	Motion JPEG
Summary	59 39	11092	3	12	4	4	2

Table 1. Variation summary of datasets used in the experiments

masks, and HKBU-MARsV1+ [19] that contains two types of masks: 6 TMF masks and 6 high-quality masks from REAL-f (RF) [4]. Both of them are recorded under controlled single lighting condition as shown in Fig. 4(a) and 4(b). To further investigate the robustness to masks with different transmittance, we also conduct experiments on the Custom Silicone Mask Attack Dataset (CSMAD) [3].

To evaluate the performance with limited observation time, we chop every long-term video into numbers of 1-second videos. As such, The 3DMAD that contains 255 10-second videos generates 2550 (255×10) samples, HKBU-MARsV1+ that contains 180 12-second videos generates 2160 (180×12) samples. CSMAD generates 1582 samples since the video length is not fixed.

HKBU-MARsV2+. To evaluate the performance under practical lighting conditions, we extend the HKBU-MARs V2 dataset [17] which contains 6 lighting conditions into a larger scale version with more variations, namely HKBU-MARsV2+. Specifically, 4 subjects and their customized REAL-f masks are added so the number of subject/mask is expanded to 16. The number of videos of each subject and mask is also increased by two times. The extended dataset is recorded at 800×600, 20fps using an industrial camera (MV-U3B). Example images of the 6 light settings (room light, dim light, bright light, warm light, side light, and down light) and the customized REAL-f masks (left four are newly added) are shown in Fig. 4(c).

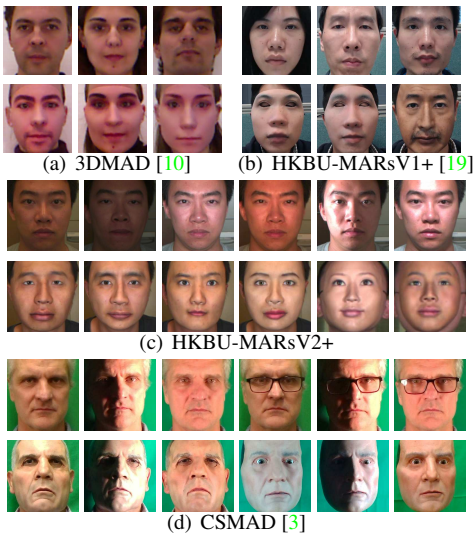
HKBU-MARsV2+ contains 480 10-second videos in total where under each lighting condition, each genuine and masked subject contains three and two videos respectively. Similarly, we chop each 10-second video into 10 1-second videos and obtain 4800 (480×10) samples in total.

Finally, we conduct extensive experiments on 11092 (2550+2160+1582+4800) video slots from the four datasets. It totally contains 39 masks from 3 mask types with different appearance quality and light transmittance, 12 lighting conditions, 4 cameras with different video quality in terms of resolution of face region and compression setting. Details of the variation we covered in this work are summarized in Tab. 1. To our best knowledge, this is the largest 3D mask dataset scale that is used to evaluate the rPPG-based PAD method.

Implementation Details. We follow [18][5] to define the 15 half-overlapped local facial ROIs (see Fig. 2). TSrPPG operator extracts the similarity of 105 possible combinations.

[4] http://real-f.jp

[5] https://github.com/boomer647/LrPPG.



(a) 3DMAD [10] (b) HKBU-MARsV1+ [19]

(c) HKBU-MARsV2+

(d) CSMAD [3]

Figure 4. Example face images from 4 3D mask face anti-spoofing datasets: (a) 3DMAD, (b) HKBU-MARsV1+, (c) the extended HKBU-MARsV2+, (d) CSMAD. Genuine and masked faces are shown in the first and seconde row, respectivly. (c) also demonstrates 6 lighting conditions and the hyper real mask examples made by REAL-f (left four are newly added).

For background TSrPPG we extract 9 larger facial regions and 6 ($J = 6$) background regions following the same definition in [19]. The shifting index n ranged from $-Fs/2$ to $Fs/2$, where Fs is the fps of the face video. The gradient feature parameter $\{L, \sigma\}$ is set as $\{7, 1\}$ empirically. Parameters of SVM are tuned automatically through MATLAB during training.

Baseline Methods. We compare our method with all existing rPPG-based 3D mask PAD methods: the global rPPG-spectrum based method GrPPG [16] and local rPPG-spectrum based method LrPPG [18], PPGSec [24], and CFrPPG [19] which is the state of the arts. We follow their original classifier settings with well-tuned parameters.

Evaluation Criteria. AUC, EER, Half Total Error Rate (HTER) [10] are used as the evaluation criterias. For intra-dataset testing, HTER on the development set (HTER_dev) and testing set (HTER_test) is measured. ROC curves with Bona fide Presentation Classification Error Rate (BPCER) and Attack Presentation Classification Error Rate (APCER) are plotted for qualitative comparisons.

5.2. Experimental Results

Intra-dataset Evaluation On the 3DMAD and HKBU-MARsV1+ with 1 second observation time, we follow the

leave one out cross validation (LOOCV) setting with random subject selection in [19, 18] for intra-dataset evaluation. Results summarized in Tab. 2 and Tab. 3 shows the strong performance of the proposed method, which outperforms the state-of-the-art rPPG-based methods in a large gap (about 20% improvement on EER and AUC). The results on 3DMAD are better than that on HKBU-MARsV1+ because the latter is recorded under uncontrolled light where the local rPPG signals are prone to being contaminated. It is worth mentioning that some subjects shows around 50-55 beats per minute heart rate as they sit quite during the dataset recording. This validates that proposed TSrPPG operator can obtain a distinguishable pattern within less one heartbeat cycle. With short observation time, the performance of CFrPPG [19], LrPPG [18] and GrPPG [16] drops dramatically compared with their reported results. This verifies our deduction that the rPPG-spectrum based approach highly depends on the length of observation time.

	HTER_dvlp	HTER_test	EER	AUC
GrPPG [16]	34.1 ± 5.7	33.7 ± 11.6	38.3	65.9
LrPPG [18]	33.3 ± 3.1	33.0 ± 8.1	34.8	69.4
PPGSec [24]	45.2 ± 3.2	44.8 ± 8.8	45.3	55.7
CFrPPG [19]	32.8 ± 1.7	32.7 ± 7.4	32.5	70.8
TSrPPG	**13.1 ± 3.0**	**13.4 ± 11.2**	**13.3**	**93.8**
TSrPPG-amp	13.4 ± 3.0	13.5 ± 11.2	13.6	93.2
TSrPPG-grdt	15.1 ± 2.8	15.4 ± 10.4	15.4	92.4
TSrPPG-phs	15.2 ± 2.9	15.2 ± 11.1	15.4	91.9
TSrPPG-no-bkgrd	14.5 ± 2.9	14.5 ± 11.1	14.6	92.4
TSrPPG-strfwd	16.2 ± 2.8	16.5 ± 10.0	16.2	91.3

Table 2. Intra dataset evaluation results on 3DMAD with short observation time (1 second)

	HTER_dvlp	HTER_test	EER	AUC
GrPPG [16]	29.2 ± 4.7	29.1 ± 9.7	33.8	72.0
LrPPG [18]	42.4 ± 2.1	42.9 ± 5.8	43.0	59.3
PPGSec [24]	45.3 ± 3.7	45.1 ± 12.0	45.3	56.2
CFrPPG [19]	41.6 ± 3.3	42.1 ± 5.6	42.0	60.8
TSrPPG	**21.5 ± 2.6**	**22.3 ± 8.8**	**22.0**	**85.2**
TSrPPG-amp	22.3 ± 3.0	22.8 ± 9.9	22.8	84.1
TSrPPG-grdt	22.3 ± 2.7	23.0 ± 8.3	23.2	84.0
TSrPPG-phs	24.0 ± 3.0	25.1 ± 9.2	24.6	82.0
TSrPPG-no-bkgrd	22.4 ± 2.7	22.9 ± 9.6	23.0	84.1
TSrPPG-strfwd	25.8 ± 2.4	26.3 ± 9.1	26.1	81.0

Table 3. Intra dataset evaluation results on HKBU-MARsV1+ with short observation time (1 second)

Ablation Study. Following the intra-dataset testing, we conduct the ablation study to evaluate the effectiveness of each TSrPPG component, TSrPPG-amp, TSrPPG-grdt, and TSrPPG-phs. The intuitive TSrPPG solution TSrPPG-strfwd is implemented and evaluated by setting $m = 0$. TSrPPG without background information (TSrPPG-no-bkgrd) is also compared. All sub-components of TSrPPG are evaluated using SVM with RBF kernel. Results in Tab. 2 and 3 illustrate the effectiveness of the each component. The comparison between TSrPPG and TSrPPG-strfwd indicates the superiority of the proposed feature operator.

We also study the performance with different length of observation time by chopping every long-term video into several half-overlapped short-term videos. Results summarized in Tab. 4 show that TSrPPG outperforms the others with different length of observation time. It is also noted that the TSrPPG with only 2-second observation outperforms the LrPPG with 10-second observation (95% AUC [18]).

	3DMAD				HKBUMARsV1+			
	1s	2s	3s	4s	1s	2s	3s	4s
GrPPG [16]	65.9	79.1	84.6	87.7	72.0	79.2	80.3	82.3
LrPPG [18]	69.4	84.1	89.3	92.0	59.3	71.5	78.8	84.5
PPGSec [24]	55.7	68.3	74.5	80.0	56.2	74.4	76.7	79.8
CFrPPG [19]	70.8	88.1	93.1	94.4	60.8	78.6	85.8	89.0
TSrPPG	**93.8**	**97.0**	**97.7**	**98.4**	**85.2**	**89.0**	**89.9**	**90.3**

Table 4. Performances (AUC) with different length of observation time.

Cross-dataset Evaluation. To evaluate the generalization ability, we follow protocols in [19] to conduct the cross-dataset experiments by training and testing with different datasets. For train on 3DMAD and test on HKBU-MARsV1+ (3DMAD→HKBUMARsV1+), the number of training subjects is 8. For HKBUMARsV1+→3DMAD, the number of training subjects is 6. All subjects in the testing dataset are used for both settings. Besides the rPPG approach, three popular appearance-based methods are added for comparison: the multi-scale LBP (MS-LBP) [22], the color texture analysis (CTA) [5], and CNN which use a pre-trained VGGNet [31] as the feature extractor (output of fc2 layer). SVM with RBF kernel with well tuned parameters is employed as the classifier.

Results summarized in Tab. 5 illustrate the robustness of the proposed method. The performance degradation of rPPG-based baselines compared with their original reported results verifies their weakness with short observation time again. It is also noted that the performance of appearance-based methods drops compared with the reported intra-dataset testing results(around 99% AUC on 3DMAD) [18], which exposes the over-fitting problem due to their data-driven property.

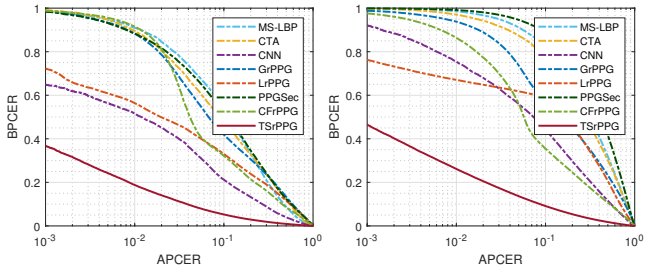
Rubustness to lighting variation. We conduct the random LOOCV (16 iterations) on HKBU-MARsV2+ where the number of training and development subjects is set as 7 and 8, after leaving 1 testing subject out. To evaluate the robustness to different lighting conditions, we conduct the leave one variation out (LOVO) [17] protocol under the LOOCV framework. Differently, the one variation left out is for training while the others are for testing. The small performance degradation of TSrPPG shown in Tab. 6 and Fig. 5 indicates its good generalizability to practical unseen lighting conditions. GrPPG and LrPPG loose the robustness with limited observation time and drop about 10% on EER and AUC. CFrPPG keeps similar results since the rPPG correspondence template can extract liveness information from noisy spetrums. It is also noted that appearance-based methods struggle to adapt the lighting variations hardly.

	3DMAD→HKBUMARsV1+			HKBUMARsV1+→3DMAD		
	HTER(%)	EER(%)	AUC(%)	HTER(%)	EER(%)	AUC(%)
MS-LBP [10]	53.0 ± 3.6	39.8	60.4	32.8 ± 11.5	32.5	75.3
CTA [4]	40.1 ± 7.8	40.2	62.1	47.7 ± 5.4	42.5	60.5
CNN	50.0 ± 0.0	47.8	54.6	50.0 ± 0.0	44.3	58.6
GrPPG [16]	46.8 ± 3.0	47.5	53.6	31.5 ± 3.8	31.1	70.0
LrPPG [18]	39.2 ± 0.8	43.1	60.1	40.4 ± 2.7	41.7	60.6
PPGSec [24]	49.7 ± 3.1	49.2	50.7	48.0 ± 2.0	47.8	52.6
CFrPPG [19]	39.2 ± 1.4	40.1	63.6	40.1 ± 2.3	40.6	62.3
TSrPPG	**23.5 ± 0.5**	**23.5**	**83.4**	**16.1 ± 1.0**	**17.1**	**90.4**

Table 5. Cross-dataset evaluation results between 3DMAD and HKBU-MARsV1+ with short observation time (1 second)

	LOOCV				LOVO for training			
	HTER_dvlp	HTER_test	EER	AUC	HTER_dvlp	HTER_test	EER	AUC
MS-LBP [10]	24.9 ± 6.8	26.0 ± 18.5	25.6	81.7	37.6 ± 6.7	38.6 ± 18.1	40.5	62.2
CTA [4]	23.3 ± 7.7	22.9 ± 18.4	26.5	81.0	35.9 ± 6.5	36.7 ± 17.1	38.9	64.4
CNN	13.5 ± 4.7	14.7 ± 13.8	15.8	91.8	21.4 ± 7.1	22.2 ± 16.1	25.1	82.4
GrPPG [16]	20.2 ± 3.5	20.4 ± 6.1	26.1	81.0	26.6 ± 5.0	26.9 ± 8.1	33.5	71.9
LrPPG [18]	22.3 ± 1.6	22.2 ± 5.3	22.5	85.0	23.6 ± 2.0	23.6 ± 5.7	33.0	74.4
PPGSec [24]	32.0 ± 2.1	32.6 ± 7.3	32.1	73.7	45.9 ± 3.9	45.6 ± 9.1	45.9	55.3
CFrPPG [19]	20.8 ± 1.6	20.7 ± 5.3	20.8	85.7	21.8 ± 1.8	21.8 ± 5.4	22.6	83.8
TSrPPG	**6.74 ± 1.0**	**6.98 ± 3.3**	**6.87**	**97.8**	**7.80 ± 1.7**	**8.08 ± 4.2**	**9.46**	**96.4**

Table 6. LOOCV and LOVO evaluation results on HKBU-MARsV2+ with short observation time (1 second)



(a) LOOCV (b) LOVO for training

Figure 5. Average ROC in log scale on HKBU-MARsV2+ using LOOCV and LOVO protocol with 1-second observation

Robustness to different masks transmittance and eyeglass occlusion. To evaluate the robustness to masks with different transmittance, we conduct experiment on CSMAD [3] that use soft silicon masks with higher transmittance. For rPPG-based method, higher mask transmittance means the rPPG signals may also be found on masked faces. CSMAD also includes other challenging variations such as the eyeglass occlusion and severe side lighting (Fig. 4(d)). Since the number of subjects and masks is different, LOOCV is not applicable. We randomly select half of the subjects for training and rest for testing and 20 rounds are conducted (named protocol 1 in Tab. 7). To evaluate the effect of eyeglass occlusion, based on protocol 1 we design another protocol (protocol 2 in Tab. 7) by removing training samples with eyeglass. We compare with the state-of-the-art rPPG based methods with 1 sec. observation. Results summed in Tab. 7 show that the proposed TSrPPG outperforms the state-of-the-art with a similar gap (around 15% AUC) as it does on the 3DMAD and HKBU-MARsV1+ dataset. The similar results achieved under protocol 1 and 2 indicates the robustness of TSrPPG to eyeglass occlusion.

	Protocol 1			Protocol 2		
	HTER_test	EER	AUC	HTER_test	EER	AUC
LrPPG [18]	41.2 ± 1.7	41.7	60.7	45.0 ± 2.9	45.1	56.4
CFrPPG [19]	35.2 ± 2.3	35.4	67.9	34.9 ± 3.7	35.1	68.4
TSrPPG	**23.0 ± 3.3**	**23.4**	**84.4**	**23.0 ± 2.2**	**23.3**	**85.5**

Table 7. Evaluation on CSMAD with short observation time (1 second)

6. Discussion and Conclusion

The observation time issue of rPPG-based 3D mask PAD is addressed in this work. We propose the TSrPPG feature operator by analyzing the local rPPG signals in time domain and introduce three types of similarity metrics based on the physical properties of rPPG. With limited observation time, our TSrPPG outperforms the state-of-the-art rPPG-based methods largely and also maintains the generalizability. The results on four 3D mask attack datasets with 25 variations on mask type, light, camera, and video quality shows the high potential of TSrPPG on real application usage.

Besides, we can find also that compared with the face resolution and lighting condition, camera with different compression settings affects more on the performance. For instance, although HKBU-MARSv1+ is recorded at high resolution compared with 3DMAD and HKBU-MARsV2+, the performances are lower (see Tab. 1). This is because the H.264 compression removes some of the subtle color variations that reflect heartbeat [28]. More self-made 3D mask attack datasets with different camera and compression settings are needed to further investigate the properties of rPPG-based 3D mask PAD.

7. Acknowledgement

This project is partially supported by Hong Kong RGC General Research Fund HKBU 12201215 and HKBU Tier 1 start-up Grant.

References

[1] A. Agarwal, R. Singh, and M. Vatsa. Face anti-spoofing using haralick features. In *BTAS*, 2016. 2

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *ICCVW*, 2013. 3

[3] S. Bhattacharjee, A. Mohammadi, and S. Marcel. Spoofing deep face recognition with custom silicone masks. In *BTAS*, 2018. 3, 6, 8

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing based on color texture analysis. In *ICIP*, 2015. 8

[5] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016. 1, 2, 7

[6] P. Buddharaju, I. Pavlidis, and C. Manohar. Face recognition beyond the visible spectrum. In *Advances in Biometrics*, pages 157–180. Springer, 2008. 5

[7] T. de Freitas Pereira, J. Komulainen, A. Anjos, J. M. De Martino, A. Hadid, M. Pietikäinen, and S. Marcel. Face liveness detection using dynamic texture. *EURASIP Journal on Image and Video Processing*, 2014(1):1–15, 2014. 1, 3

[8] G. de Haan and V. Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 3, 5

[9] N. Erdogmus and S. Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *BTAS*, 2013. 3, 5

[10] N. Erdogmus and S. Marcel. Spoofing face recognition with 3d masks. *IEEE Transactions on Information Forensics and Security*, 9(7):1084–1097, 2014. 1, 2, 3, 6, 8

[11] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE transactions on image processing*, 23(2):710–724, 2014. 1, 2

[12] J. Guo, X. Zhu, J. Xiao, Z. Lei, G. Wan, and S. Z. Li. Improving face anti-spoofing by 3d virtual synthesis. *arXiv*, 2019. 2

[13] A. Jourabloo, Y. Liu, and X. Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV*, pages 290–306, 2018. 2

[14] G. Lempe, S. Zaunseder, T. Wirthgen, S. Zipser, and H. Malberg. Roi selection for remote photoplethysmography. In *Bildverarbeitung für die Medizin 2013*, pages 99–103. Springer, 2013. 5

[15] X. Li, J. Chen, G. Zhao, and M. Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *CVPR*, 2014. 3

[16] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen. Generalized face anti-spoofing by detecting pulse from face videos. In *ICPR*, 2016. 2, 3, 6, 7, 8

[17] S. Liu, B. Yang, P. C. Yuen, and G. Zhao. A 3d mask face anti-spoofing database with real world variations. In *CVPRW*, 2016. 2, 6, 7

[18] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *ECCV*, 2016. 2, 3, 4, 6, 7, 8

[19] S.-Q. Liu, X. Lan, and P. C. Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In *ECCV*, pages 558–573, 2018. 2, 3, 6, 7, 8

[20] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398, 2018. 2

[21] 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. *arXiv*, 2019. 1, 2

[22] J. Määttä, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IJCB*, 2011. 1, 2, 7

[23] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão, and A. Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4):864–879, 2015. 2

[24] E. M. Nowara, A. Sabharwal, and A. Veeraraghavan. Ppgsecure: Biometric presentation attack detection using photopletysmograms. In *FG*, 2017. 3, 6, 7, 8

[25] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, 2007. 1, 3

[26] K. Patel, H. Han, A. K. Jain, and G. Ott. Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks. In *ICB*, 2015. 2

[27] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 3

[28] M. Rapczynski, P. Werner, and A. Al-Hamadi. Effects of video encoding on camera based heart rate estimation. *IEEE Transactions on Biomedical Engineering*, 2019. 8

[29] D. Shao, Y. Yang, C. Liu, F. Tsow, H. Yu, and N. Tao. Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time. *IEEE Transactions on Biomedical Engineering*, 61(11):2760–2767, 2014. 3

[30] K. Shelley and S. Shelley. Pulse oximeter waveform: photoelectric plethysmography. *Clinical Monitoring, Carol Lake, R. Hines, and C. Blitt, Eds.: WB Saunders Company*, pages 420–428, 2001. 3

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 7

[32] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 1, 2

[33] J. Yang, Z. Lei, and S. Z. Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 2

[34] D. Yi, Z. Lei, Z. Zhang, and S. Z. Li. Face anti-spoofing: Multi-spectral approach. In *Handbook of Biometric Anti-Spoofing*, pages 83–102. Springer, 2014. 1