3D Face Anti-Spoofing with Factorized Bilinear Coding

Shan Jia, Xin Li, *Fellow, IEEE,* Chuanbo Hu, Guodong Guo, *Senior member, IEEE,* and Zhengquan Xu

Abstract—We have witnessed rapid advances in both face presentation attack models and presentation attack detection (PAD) in recent years. When compared with widely studied 2D face presentation attacks, 3D face spoofing attacks are more challenging because face recognition systems are more easily confused by the 3D characteristics of materials similar to real faces. In this work, we tackle the problem of detecting these realistic 3D face presentation attacks and propose a novel anti-spoofing method from the perspective of fine-grained classification. Our method, based on factorized bilinear coding of multiple color channels (namely MC_FBC), targets at learning subtle fine-grained differences between real and fake images. By extracting discriminative and fusing complementary information from RGB and YCbCr spaces, we have developed a principled solution to 3D face spoofing detection. A large-scale wax figure face database (WFFD) with both images and videos has also been collected as super realistic attacks to facilitate the study of 3D face presentation attack detection. Extensive experimental results show that our proposed method achieves the state-of-the-art performance on both our own WFFD and other face spoofing databases under various intra-database and inter-database testing scenarios.

Index Terms—Face spoofing attack, presentation attack detection, wax figure face, face anti-spoofing.

I. INTRODUCTION

FACE has been one of the most widely used biometric modalities due to its accuracy and convenience for personal verification and identification. However, the increasing popularity and easy accessibility of face modalities make face recognition systems (FRS) a major target of spoofing such as presentation attacks [1]. This class of security threats can be easily implemented by presenting the FRS a face artifact, which is also known as presentation attack instrument (PAI) [2]. A recent breach of biometrics database (BioStar) leads to the compromise of as many as 28 million records containing facial and fingerprint data, which can be exploited by malicious hackers as PAIs.

Based on the way of generating face artifacts, face presentation attacks can be classified into 2D (e.g., printed/digital photographs or recorded videos on mobile devices such as a tablet) and 3D (e.g., by wearing a mask or presenting a synthetic model). Existing research on FRS has paid more attention to 2D face PAI due to its simplicity, efficiency, and low cost. However, as material science and 3D printing technology advance, creating face-like 3D structures or materials
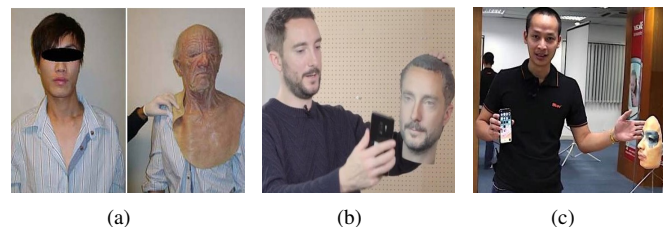

(a)   (b)   (c)

Fig. 1. Examples of 3D presentation attack cases. (a) Airport security system fooled by silicon mask[1], (b) Android phones fooled by a 3D-printed head[2], (c) iPhone X face ID unlocked by a 3D mask[3].

has become easier and more affordable. When compared to 2D attacks, 3D face presentation attacks are more realistic and therefore more difficult to be detected. The class of 3D face presentation attacks includes wearing wearable facial masks [3], building 3D facial models [4], through facial make-up [5], [6], and using plastic surgery. Fig. 1 shows several examples of 3D presentation attacks, which have successfully fooled some widely used FRS in practice.

The vulnerability of current face recognition systems to realistic face presentation attacks has facilitated a series of studies on 3D face presentation attack detection (PAD) [7]. Existing methods tried to explore the difference between real face skin and 3D fake face materials based on the reflectance properties using multispectral imaging [8], [9], texture analysis [10], [11], deep features [12], [13], or liveness cues [14], [15]. They have achieved promising detection performance on several existing 3D face spoofing databases [3], [16], [17], [18]. However, since facial masks are often made of paper, latex or silicone materials, these 3D face spoofing databases have the limitations of small database size (mostly less than 30 subjects), poor authenticity (some based on 2D planar or noncustomized masks [17], [19]), and low diversity in subject and recording process, which might impede the development of effective and practical PAD schemes. Several studies [7], [20], [21], [22] have shown that a variety of 3D PAD methods suffer from performance degradation on databases with more diverse and realistic 3D face spoofing attacks.

In this work, we aim at detecting realistic 3D face spoofing attacks which often pose greater threats to existing FRS (even with face anti-spoofing modules) than their 2D counterparts. Based on the fact that wax figure faces have already been used for identity, personation, and fraud in the real world

S. Jia and Z. Xu are with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: jias@whu.edu.cn; xuzq@whu.edu.cn).

X. Li, G. Guo, and C. Hu are with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA (e-mail: {xin.li, chuanbo.hu, guodong.guo}@mail.wvu.edu)

Corresponding author: Xin Li.

[1]Picture is downloaded from https://chameleonassociates.com/security-breach/.

[2]Picture is downloaded from http://www.floridaforensicscience.com/broke-bunch-android-phones-3d-printed-head/.

[3]Picture is downloaded from https://boingboing.net/2010/11/05/young-asian-refugee.html.

Fig. 2. Photos with wax figure faces for fraud. In 2012, six suspects snapped about 600,000 people out of nearly US$475 million under a pyramid sales scam using photos taken with the wax figures at Hong Kong's Madam Tussauds Museum, (a) with the wax figure of chief executive Donald Tsang Yam-Kuen, (b) with the wax figure of business tycoon Li Ka-Shing.

(as shown in Fig. 2), we first collect a large number of wax figure faces based on both still images and moving videos as super realistic 3D face spoofing attacks by expanding upon our preliminary work [23]. Furthermore, we propose to treat realistic face spoofing detection as a special class of fine-grained image classification problems which focus on differentiating between hard-to-distinguish object classes. Then we introduce a new bilinear coding based method with state-of-the-art performance of combining features for fine-grained classification, which generates discriminative representations for 3D face anti-spoofing. The main contributions of this work are summarized below.

- A large-scale wax figure face database (WFFD) with both images and videos is constructed as the super realistic 3D face presentation attack. It contains 2300 pairs of matched faces from 745 subjects (totally 4600 faces), and 285 videos from 241 subjects (with over 45000 frames).
- A new 3D face anti-spoofing method, named MC_FBC, is proposed based on factorized bilinear coding for multicolor channels. This is the first work addressing the problem of detecting a realistic face PAD in a fine-grained manner and use bilinear coding to improve the discriminative power of feature representations.
- We have conducted extensive experiments on the proposed WFFD and several publicly available databases. Our findings show that the proposed method outperforms several state-of-the-art methods under both intra- and inter-database testing scenarios.

The rest of this paper is organized as follows. In Section II, we briefly review related research in 3D face PAD methods and spoofing databases. Section III introduces the WFFD database with both photo-based and video-based wax figure faces as super realistic 3D face spoofing attacks. The proposed 3D face anti-spoofing method based on factorized bilinear coding is presented in Section IV. Experimental results are reported in Section V, and we make several conclusions about this work and future research in Section VI.

II. RELATED WORK

A. 3D Face PAD methods

Detection of 3D fake faces is often more challenging than detecting fake faces with 2D planar surfaces. Methods designed for 2D face spoofing detection may fail to identify 3D face spoofing attacks, especially for those using recapture effects [24] and spoofing medium contour detection [25]. Existing PAD methods for 3D face presentation attacks, mainly based on the difference between real face skin and mask material, can be broadly classified into five categories, namely, reflectance-based, texture-based, shape-based, liveness-based, and deep feature based.

Earlier studies [26], [27], [28] in 3D mask spoofing detection were based on the reflectance difference of facial skin and mask materials. For example, the distribution of albedo values for illumination of various wavelengths was first analyzed in [26] to find how different facial skins and mask materials (silicon, latex, and skinjell) behave in terms of reflectance. Texture-based methods explore the texture pattern difference of real faces and masks with the help of texture feature descriptors, such as the widely used Local Binary Patterns (LBP) [3], [29], Binarized Statistical Image Features (BSIF) [30], and Haralick features [10].

Shape-based 3D mask PAD methods use shape descriptors [31], [32], [33] or 3D reconstruction [11] to extract discriminative features from faces and 3D masks. Different from reflectance-based or texture-based detection methods, these schemes only require standard color images without the need of special sensors. However, their detection performances rely on the quality of 3D mask attacks, and may not be robust to super realistic 3D face presentation attacks. More recently, some methods explore liveness cues to detect 3D face presentation attacks, such as thermal signatures [34], gaze information [35], [36], and pulse or heartbeat signals [20], [37], [38], [39]. Based on the intrinsic liveness signals, these methods achieve outstanding performance on distinguishing real faces from masks.

Instead of extracting hand-crafted features, deep feature based methods automatically extract features from face images. Two deep representation approaches were investigated in [40] for spoofing detection in different biometric modalities. Image quality cues (Shearlet) and motion cues (dense optical flow) were fused in [41] using a hierarchical neural network for mask spoofing detection, which achieved a Half Total Error Rate (HTER) of 0% on the 3DMAD database [3]. A network based on transfer learning using a pre-trained VGG-16 model architecture is presented in [42] to recognize photo, video, and 3D mask attacks. Based on the observation of the importance of dynamic facial texture information, a deep convolutional neural network-based approach was developed in [12], [21]. Both intra-dataset and cross-dataset evaluation on 3DMAD and their supplementary dataset indicated the efficiency and robustness of the proposed method.

Despite these advances in 3D face anti-spoofing, there are limitations in each category of detection methods. For example, the main limitation of reflectance-based methods is the requirement of special and expensive devices to acquire multispectral images at varying wavelengths. Although texture and shape-based methods are easy-to-implement, their robustness to different mask spoofing attacks needs further investigation. The liveness cues, such as pulse/heartbeat-based, are highly dependent on lighting conditions and camera settings (e.g., exposure time and frame rates). Deep feature based approaches are generally sensitive to dataset sizes and lack transparency.

In addition, most of them [7], [20], [43], [21], [22], [44], [35], [37], [38] still suffer from performance degradation when applied to databases with more realistic face spoofing attacks. How to improve the *robustness* and *discriminative power* of representations to distinguish real faces from fake ones has remained a long-standing open problem.

B. 3D Face Spoofing Databases

Existing 3D face spoofing databases create attacks mainly based on wearable 3D face masks that are made of material with face characteristics similar to real faces. 3DMAD [3] is the first publicly available 3D mask database. It used the services of ThatsMyFace[4] to manufacture 17 masks of users, and recorded 255 video sequences with an RGB-D camera of Microsoft Kinect device for both real access and presentation attacks. With the development of 3D modeling and printing technologies, more mask databases have been created since 2016. 3DFS-DB [45] is a self-manufactured and gender-balanced 3D face spoofing database, in which 26 printed models were made using two 3D printers. HKBU-MARs [16], [20] is another 3D mask spoofing database with more variations. It generated customized masks from two companies (ThatsMyFace and REAL-F[5]), and created videos from 12 subjects. To include more subjects, SMAD database [17] has collected videos of people wearing silicone masks from online resources. It contains 65 genuine videos of people auditioning, interviewing, or hosting shows, and 65 imposter videos of people wearing a complete 3D (but not customized) mask which fits well with proper holes for the eyes and mouth.

For effective detection, more 3D face spoofing databases provide multi-modality spoofing samples under special lighting conditions. The BRSU Skin/Face/Spoof Database [28] provides multispectral SWIR (Short Wave Infrared) and RGB color images for various types of masks and facial disguises. The MLFP database [19] is another multispectral database using latex and paper masks. It contains 1350 videos of 10 subjects in visible, near infrared (NIR), and thermal spectrums. The ERPA database [34] provides RGB and NIR images of both bona fide and 3D mask attack presentations captured using special cameras. This is a small dataset with only 5 subjects involved. Both rigid resin-coated masks and flexible silicone masks are considered. Similarly, the recently released WMCA database [18] used multiple capturing devices/channels, including color, depth, thermal, and infrared. It contains 1679 videos with 347 bona fide and 1332 attacks from 72 subjects; among them, 709 3D face spoofing attack videos include fake head, rigid mask, flexible silicone masks, and paper masks. The 3DMA [46] also collects 920 videos captured in both visual and NIR modalities from 67 subjects.

These databases have played a significant role in designing multiple detection schemes against 3D face presentation attacks. However, the following issues have remained to be addressed: (1) only a small number of 3D spoofing samples are collected due to the difficulty and expense of mask production;
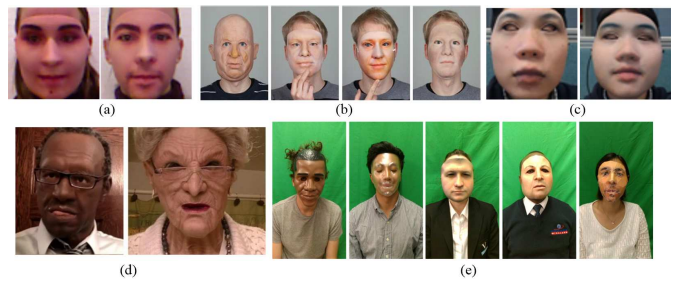

Fig. 3. 3D mask spoofing examples in existing databases. (a) 3DMAD, (b) 3DFS-DB, (c) HKBU-MARs, (d) SMAD, (e) WMCA.

(2) most of the 3D face spoofing attacks are of low quality (e.g., using noncustomized masks [17], [19] or 2D cut-out paper masks [19], [47]. Although soft/flexible masks are closer to real faces, they are still easy to be recognized by humans [3], [16], [18], [45] (as shown in Fig. 3); (3) they contain less variations to simulate the real world scenarios in terms of recording cameras, lighting settings, subject pose/age, and facial expression/resolution. Therefore, we have introduced wax figure faces as super realistic 3D face spoofing attacks in [23] (with 4400 still faces). After that, Vareto et al. [48] collected a wax figure database named SWAX with 1800 faces and 110 videos from 55 people/waxworks.

In this paper, to overcome the limitations in both 3D face spoofing databases and detection methods, we first expand the wax figure face dataset to include both images and videos with a high degree of diversity, aiming at more faithfully simulating realistic spoofing attacks. Then we will design a novel detection method based on the fusion of complementary skin-inspired features in different color spaces to extract subtle differences between real faces and realistic spoofing attacks.

III. WAX FIGURE FACE DATABASE

In this section, we introduce the super realistic Wax Figure Face Database (WFFD) with both photo-based and video-based wax figure faces to address the weaknesses in the existing 3D face spoofing databases. Considering the difficulty and expense in generating a large number of 3D wax figure faces, we take advantage of the popularity and publicity of numerous celebrity wax figure museums in the world, and collect as many wax figure faces as possible from online resources to construct such a database with a large size and high diversity. These wax figure faces are all carefully designed and made in clay with wax layers, silicone, or resin materials, so that they are super realistic and similar to real faces. In the physical domain, the attacker may fool the FRS using just the life-size wax heads, which can be obtained with relatively low cost (about 200 dollars) from online shopping websites (such as eBay US and Taobao China).

A. Wax figure face database of still images

Our previous work [23] has been greatly expanded to qualify wax figure faces as super realistic 3D face spoofing attacks. We have thoroughly cleaned the dataset and removed images with tiny faces or low quality due to lossy compression, and then added 100 new pairs of images from 75 subjects to the new database. Finally, a total of 2300 pairs of images (4600

---

[4]http://thatsmyface.com/
[5]http://real-f.jp/en_the-realface.html

Fig. 4. Examples in the WFFD database. (a) in Protocol I, (b) in Protocol II. Note that one subject may have several wax figures in each protocol.

faces with both real and wax figure faces) of 745 subjects are collected. We have followed the three protocols designed in [23]: Protocol I with wax figure faces and real faces grouped manually from different devices and environmental conditions (e.g., Fig. 4(a)); Protocol II with wax figure and real faces recorded in the same environment with the same camera (e.g., Fig. 4(b)); and Protocol III combining the previous two protocols to simulate real-world operational scenarios. Table I provides the details about the statistics of images, faces, and subjects in each protocol.

TABLE I
DETAILS OF EACH PROTOCOL IN THE WFFD

| Protocol | #Image | | | | #Face | #Subject |
	Train	Valid	Test	Total		
Protocol I	600	200	200	1000	2000	462
Protocol II	780	260	260	1300	2600	409
Protocol III	1380	460	460	2300	4600	745

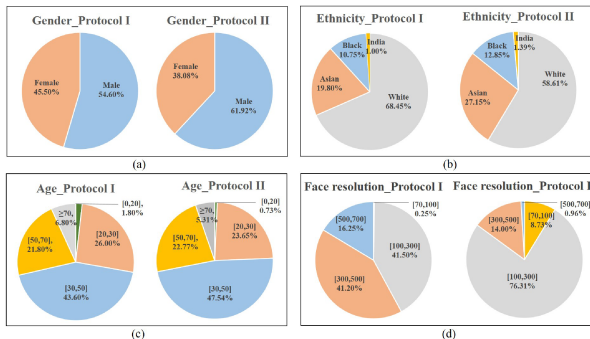Note that the train, validation, and test subsets are non-overlapped.

Fig. 5. Statistical distribution of the WFFD. (a) Gender, (b) ethnicity, (c) age, (d) face resolution.

The statistical information about subject gender, age, ethnicity (detected by Face++[6]), and face resolution (cropped by the dlib face detector [49]) in the WFFD is shown in Fig. 5. It can be seen that the images in the WFFD are relatively gender balanced - with about 60% of males and 40% of females in both protocols. The ethnicity distribution in Fig. 5(b) contains a majority of White subjects (around 60%), followed by about 20% Asians and 10% Blacks, and a small percentage of Indians (no more than 2%). We can also see a wide distribution of ages in Fig. 5(c). The two protocols have similar distribution patterns in terms of age, with half subjects being between 30 and 50 years old. Although the dimensions

[6]https://www.faceplusplus.com/face-compare-sdk/

of most face regions are between 100×100 and 500×500, there is a big difference in the distribution between the first two protocols. Matched and grouped manually, the dimensions of face regions in Protocol I are generally larger than those in Protocol II. Additionally, the images in Protocol I are more diversified in terms of subject poses, facial expression, recording environment, and devices than those in Protocol II.

B. Wax figure face database of videos

Inspired by the new generation of intelligent or robotic wax figures which can move and interact with visitors (like the one unveiled by Madame Tussauds in Shanghai in 2018), we have further collected video-based wax figure faces from online resources to obtain moving wax figure faces. Similar to the collection of still wax figure faces, we first download as many short videos with celebrity wax figure faces as possible. Then we clean the dataset manually based on the selection criterion that videos without frontal faces, or with faces but containing a significant amount of occlusion or embedded text, should be excluded from the dataset [50]. Finally, a total of 145 wax figure videos, as well as 140 real face videos have been collected as the newly-constructed WFFD-V database. All real and fake videos are between 60 and 420 frames in length. More details of the WFFD-V with 285 gender-balanced videos are included in Table II. They are randomly partitioned into non-overlapped training, validation, and testing subsets in the same ratio of 3:1:1 as the WFFD still image dataset for performance evaluation.

TABLE II
DETAILED CHARACTERISTICS OF NEWLY CONSTRUCTED WFFD-V DATASET.

| Video | #Subject | | | #Video | | | #Frame |
	F	M	Total	F	M	Total	
Real	53	74	**127**	58	82	**140**	28970
Wax	47	76	**123**	61	84	**145**	16765
Total	96	145	**241**	119	166	**285**	45735

F denotes Female, and M denotes Male.

IV. 3D FACE ANTI-SPOOFING BASED ON FACTORIZED BILINEAR CODING

With materials and shapes highly similar to real faces, 3D face spoofing attacks often lead to performance degradation of existing face PAD methods [29]. To detect realistic 3D spoofing attacks, we propose to explore their subtle differences from real faces based on generating discriminative features in a fine-grained manner. Inspired by recent advances in fine-grained classification [51], [52], [53], [54], and subtle variation detection [55], we propose to tackle the problem of realistic face spoofing detection by combining the skin color model [56] with factorized bilinear pooling [57]. Originally proposed for fine-grained visual recognition, bilinear pooling [58] has become a popular tool for multimodal data fusion due to its superiority in exploiting higher-order information among complementary features. In this section, we first construct complementary skin-inspired features in color space and then present a novel method of factorized bilinear coding combining features extracted in different color spaces.

A. Skin color model

Color spaces play an important role in image processing and computer vision applications. RGB is the most widely used color space for sensing, representing, and displaying color images. However, due to the high correlation among the three color components (red, green, and blue), RGB color space representation is not necessarily the most appropriate choice for face anti-spoofing. Alternative color-space representations such as luminance/chrominance and hue/saturation are also valid and competing choices. Instead of fusing features from a single RGB color space as most previous bilinear pooling schemes [57], [58], [59], [60], [54] did, we propose to take multiple color spaces into account and extract more discriminative features by combining multicolor space representations for 3D face anti-spoofing.

Different from RGB color space, YCbCr color space encodes a color image similar to human eyes' retina, which separates the RGB components into a luminance component (Y) and two chrominance components (Cb as blue projection and Cr as red projection; for an analog version, as U and V, respectively). YCbCr space is effective for color feature extraction and has also achieved promising performance in face-related applications (e.g., human face detection [56], 2D face spoofing detection [61], and skin classification [62]).

Considering the high similarity between 3D face spoofing attacks and real faces, we propose to take both RGB and YCbCr color spaces of the face image as the input of CNN-based feature extraction module to obtain robust facial color texture descriptions. The key motivation behind combining different color textures is two-fold. First, each *color-texture* based analysis can help capture the artifacts of spoofing attacks for detection. The artificial face in face spoofing attacks is either made of special materials different from human skins (e.g., silicon, latex, and skinjell mask) or passes through different camera systems or printing systems (or display devices in 2D attacks) [63]. Therefore, artificial face images are likely to suffer from different kinds of quality degradation issues - e.g., the face production material artifacts, the PAI dependent color variations, and limited color reproduction in 2D print or replay spoofing, which do not occur in real faces. Second, the fusion of two different color spaces, namely, the RGB with highly correlated color components, and YCbCr with separated luminance and chrominance components, has the potential of learning *complementary* and robust subtle features for face anti-spoofing (see the results in Section V-B and V-G).

B. Factorized bilinear coding

Bilinear pooling was introduced in [58] to provide robust image representation for fine-grained image classification. In bilinear pooling models, two feature vectors are fused by an outer product (or Kroneker product of matrices); this way, all pairwise interactions among the given features are considered as follows:

$$Z = \sum_{(i,j)\in\mathbb{S}} x_i y_j^\top \qquad (1)$$

where $\{x_i | x_i \in \mathbb{R}^p, i \in \mathbb{S}\}$, $\{y_j | y_j \in \mathbb{R}^q, j \in \mathbb{S}\}$ are two input features, $\mathbb{S}$ is the set of spatial locations (combinations of

rows and columns), and $Z \in \mathbb{R}^{p\times q}$ is the fused feature descriptor. It can be seen that the size of the bilinear feature descriptor can be large, which makes it computationally infeasible. To generate more compact representations, we have employed the factorized bilinear coding (FBC) [57] to more computationally efficiently integrate the features from both RGB and YCbCr color spaces for 3D face anti-spoofing.

Let x_i, y_j be the two features extracted from RGB and YCbCr color spaces respectively, the FBC encodes the features based on sparse coding, and learns a dictionary B with k atoms that are factorized into low-rank matrices to capture the structure of the whole data space. Specifically, let the dictionary $B = [b_1, b_2, ..., b_k] \in \mathbb{R}^{pq\times k}$, and FBC proposes to factorize each dictionary atom $b_l \in \mathbb{R}^{pq}(1 \leq l \leq k)$ into $U_l V_l^\top$, where $U_l \in \mathbb{R}^{p\times r}$ and $V_l \in \mathbb{R}^{q\times r}$ are learnable low-rank matrices (as the hyper-parameter rank $r \ll p, q$). Therefore, the original bilinear features $x_i y_j^\top$ can be reconstructed by $\sum_{l=1}^k c_v^l U_l V_l^\top$, with $c_v \in \mathbb{R}^k$ being the FBC code, and c_v^l being the l-th element of c_v $(1 \leq v \leq N, N$ is the number of pairs in $\mathbb{S}$). Then the sparsity-based FBC encodes the two input features (x_i, y_j) into c_v by solving the following optimization problem,

$$\min_{c_v} \left\| x_i y_j^\top - \sum_{l=1}^k c_v^l U_l V_l^\top \right\|^2 + \lambda \|c_v\|_1 \qquad (2)$$

where λ is a trade-off parameter between the reconstruction error and the sparsity. To obtain the FBC code c_v, the classical LASSO method [64] has been adopted as shown in Eq. (3).

$$\begin{cases} c_v' = P(\widetilde{U}^\top x_i \circ \widetilde{V}^\top y_j), \\ c_v = \text{sign}(c_v') \circ \max((\text{abs}(c_v') - \frac{\lambda}{2}), 0). \end{cases} \qquad (3)$$

where $\circ$ denotes the Hadamard product, $P \in \mathbb{R}^{k\times rk}$ is a fixed binary matrix with only elements in the row l, columns $((l-1)r+1)$ to (lr) being "1", and $\widetilde{U}$ and $\widetilde{V}$ are the transformations of U and V to avoid matrix inversion with heavy computation in the original LASSO method. They are in the form of

$$\begin{cases} \widetilde{U}^\top = [\widetilde{U}_l^\top] = [\frac{1}{r} \cdot I((q_l 1_{rk}^\top) \circ U^\top)] \in \mathbb{R}^{rk\times p} \\ \widetilde{V}^\top = [\widetilde{V}_l^\top] = [\frac{1}{r} \cdot I V^\top] \in \mathbb{R}^{rk\times q} \end{cases} \qquad (4)$$

where $I \in \mathbb{R}^{r\times rk}$ is an all "1" matrix, q_l is the l-th column of $Q = ((P(U^\top U P^\top \cdot V^\top V P^\top))^{-1} P)^\top$.

With Eq. (3), the FBC code c_v can be obtained by learning $\widetilde{U}$ and $\widetilde{V}$ instead of U and V. We can get all FBC codes c in feature pair set $\mathbb{S}$; then they are fused using the max operation to attain the final global representation $z = max\{c_v\}_{i=1}^N$. The FBC module is applied to the features extracted by the last convolutional layer of a CNN (e.g., VGG), then followed by a fully connected (FC) layer for classification using Softmax classifier. The whole process of the proposed MC_FBC scheme is shown in Fig. 6.

C. Loss Function

To train the network, we utilize the focal loss function [65] which reshapes the standard cross entropy loss in such a way that the loss assigned to well-classified examples in binary
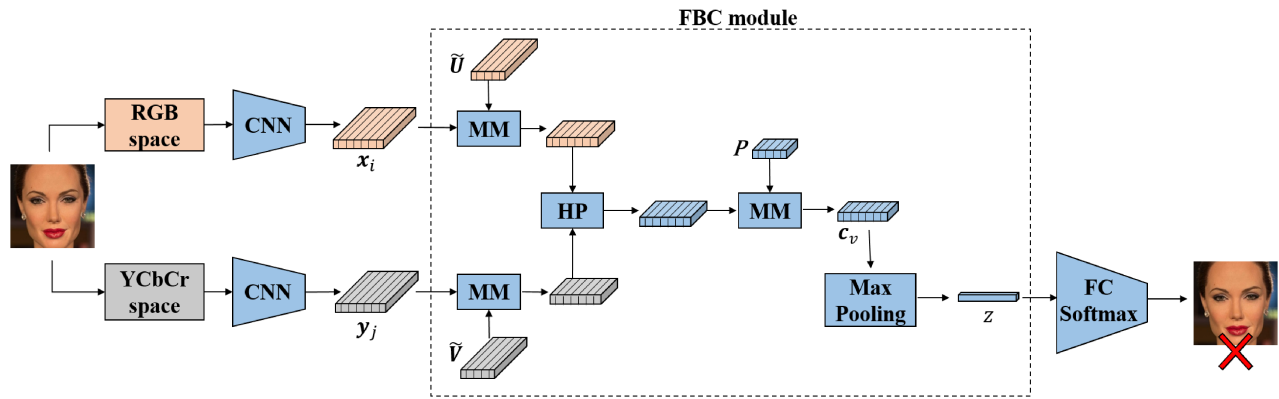
Fig. 6. Flowchart of the proposed MC_FBC scheme for 3D face anti-spoofing. The FBC module generates discriminative representations in a fine-grained manner. MM module refers to 'Matrix Multiplication', and HP module refers to 'Hadamard Product' operation. The variables x_i and y_j are the features learned from CNN, $\widetilde{U}$ and $\widetilde{V}$ are learnable parameters, P is a fixed binary matrix with only elements in specific rows and columns being "1", c_v is the FBC code, and Z is the final global representation.

classification is downweighted. There are two compelling reasons for such choice of loss function: 1) the higher similarity of real faces and realistic 3D spoofing attacks in 3D face spoofing will lead to harder examples with large errors while the local loss function focuses on training on hard negatives and reducing the loss contribution from easy examples; 2) most 3D face spoofing databases have the problem of class imbalance (e.g., more real samples than fake ones due to the difficulty of collecting 3D spoofing attacks at a large scale).

Overall, our scheme is different from [57] in the following two aspects. First, to get more discriminative color features, we extract features from two different complementary color spaces. Such diversity in terms of feature representation is important to 3D face anti-spoofing, which is sometimes even challenging for human-based inspection. Second, we replace the original cross-entropy loss function with the robust focal loss function to train the network. Originally designed for dense object detection in [65], we have found that in the scenario of anti-spoofing where training on a sparse set of hard examples is common, it is important to prevent the majority of easy negatives from dominating the detector. Such observation contributes to the improved training performance of new loss functions over the traditional ones.

V. EXPERIMENTAL RESULTS

In this section, we conduct extensive experiments to evaluate our method. We first explore the influence of color spaces on detection performance using the super realistic 3D face spoofing database. Different feature fusion schemes are also compared with our MC_FBC method. Then we present the comparison results under intra-database testing on the WFFD database and several 2D/3D face spoofing databases, and finally the performance under inter-database testing is evaluated to show the generalizability of the proposed method.

A. Experimental settings

1) Implementation: Two backbone networks pretrained on ImageNet [66] were fine-tuned on face spoofing datasets to extract features for the FBC module, namely a relatively small VGG-16 model [67] and ResNet-50 [68] (deeper and more accurate). It is worth noting that the MC_FBC can be also

applied to other network structures for further improvement of detection performance. The reasons why we used VGG-16 and ResNet-50 models are not only for their robust performance in different detection tasks, but also for a fair comparison with other bilinear pooling based methods where the two models, especially the VGG-16, are widely used [57], [58], [59], [54]. The last pooling layer and the fully-connected layers were removed in both networks. The learnable parameters $\widetilde{U}$ and $\widetilde{V}$ were updated by the back-propagation algorithm to get the FBC code. We set the rank of $\widetilde{U}$ and $\widetilde{V}$ as one, the number of dictionary atoms k as 2048, and λ as 0.001 (according to [57]). As for the training parameters, we fine-tuned the model using SGD with an initial learning rate of 0.01 (decreased by a factor of 10 for every 40 epochs until it reaches 0.0001), weight decay as 5×10^{-4}, momentum as 0.9, and batch size as 16 for VGG-16 model and 8 for ResNet-50 model. The weighting factor and tune-able focusing parameter in the focal loss function were all set to one. Input images are the cropped faces based on the *dlib* face detector [49] with size of 224×224. All experiments are conducted using PyTorch on a workstation with Titan XP GPUs.

2) Databases: In addition to our WFFD database, we have used two publicly available 3D face spoofing databases - namely, 3DMAD [3] (the most widely-used) and HKBU-MARs-V1+ [20] (with hyper-real 3D masks). As both databases contain videos of 300 frames (3DMAD with 255 videos and HKBU-MARs-V1+ with 180), 20 frames were randomly selected for spoofing detection. The averaged scores of these frames were computed as the final score. We have followed the leave-one-out cross-validation (LOOCV) protocol settings for the two databases as previous works did [12], [20], [37]. We have conducted 20 rounds of LOOCV with each round randomly selecting 10 subjects for training and 6 for validation on 3DMAD, while 5 for training and 6 for validation on HKBU-MARs-V1+ database. To validate the generalization property, we also considered two 2D face spoofing databases with both printed photo attacks and replayed video attacks: MSU-USSA [24] and Oulu-NPU database [69]. MSU-USSA database was specifically created to simulate face spoofing attacks with diversities of environment, image quality, and acquisition device. It consists of 9,000 images

TABLE III
COMPARISON RESULTS (%) OF DIFFERENT COLOR SPACES ON WFFD DATABASE UNDER PROTOCOL III

Color space	Backbone-VGG-16				Backbone-ResNet-50			
	Accuracy	APCER	BPCER	ACER	Accuracy	APCER	BPCER	ACER
RGB	93.26	8.91	4.57	6.74	93.59	8.91	3.91	6.41
YCbCr	92.39	8.26	6.96	7.61	93.91	8.70	**3.48**	6.09
YUV	87.83	11.09	13.26	12.18	92.93	9.78	4.35	7.07
HSV	87.07	16.30	9.57	12.94	91.74	10.65	5.87	8.26
YUV+YCbCr	91.09	10.87	6.96	8.92	93.37	5.87	7.39	6.63
RGB+HSV	93.15	8.26	5.43	6.85	93.91	6.30	5.87	6.09
RGB+YUV	94.24	7.17	**4.35**	5.76	94.89	5.87	4.35	5.11
RGB+YCbCr (MC_FBC)	**94.57**	**6.09**	4.78	**5.44**	**95.22**	**5.65**	3.91	**4.78**

(1,000 live subjects and 8,000 spoofing attacks) recorded with two types of cameras. A five-fold subject-exclusive cross-validation protocol was designed for this database. Oulu-NPU database contains 4,950 videos of 55 subjects with both real access and 2D face spoofing attacks. The videos were recorded using six mobile devices in three sessions with different illumination conditions, and they were divided into three subsets for training, validation, and testing, with four protocols.

3) Evaluation metrics: We report all experimental results following the ISO/IEC 30107-3 metrics [70]. Three types of errors, i.e., Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) are used in addition to the detection accuracy.

B. Ablation Study

1) The impact of multiple color spaces: We first demonstrate the effectiveness of combining multiple color spaces in detecting wax figure faces from real ones on the WFFD dataset. Table III shows the comparison results under Protocol III with two features both from the same color space (including RGB, YCbCr, YUV, and HSV) and from two different color spaces as the input of FBC module. It can be observed that in single color space, RGB and YCbCr get higher classification accuracy and lower error rates than YUV and HSV color space. However, using multiple color spaces, especially combining the RGB with YCbCr, obtains better results. Specifically, the proposed MC_FBC scheme combining RGB and YCbCr color spaces achieved 94.57% accuracy and 5.44% ACER under the VGG-16 model, while the performance was further improved by the deeper ResNet-50 model, with the highest accuracy of 95.22% and the lowest ACER of 4.78%. This shows the complementary properties of RGB and YCbCr spaces in generating more discriminative representations.

2) Exploring different feature fusion schemes: We next compare the proposed MC_FBC scheme with several feature fusion schemes, including concatenation (simply combining two feature vectors by concatenating them), score-level fusion of two color spaces, traditional bilinear pooling (BP) [58], compact bilinear pooling (CBP) [59], and factorized bilinear coding (FBC) [57]. As all these bilinear pooling based methods fused features from VGG-16 model, we present the comparison results in Table IV using VGG-16 network as the backbone. We can see that without bilinear pooling

TABLE IV
COMPARISON RESULTS (%) OF DIFFERENT FUSION METHODS ON WFFD DATABASE UNDER PROTOCOL III USING VGG-16

Fusion scheme	Accuracy	APCER	BPCER	ACER
Original VGG-16	84.68	13.48	17.17	15.33
Concatenation	84.35	13.91	17.39	15.65
Max score fusion	85.43	12.39	16.74	14.56
Mean score fusion	85.87	11.96	16.30	14.13
BP [58]	92.28	7.61	7.83	7.72
CBP [59]	90.87	8.48	9.78	9.13
FBC [57]	92.83	8.70	5.65	7.18
FBC_FL	93.26	8.91	**4.57**	6.74
MC_FBC_CE	93.91	7.39	4.79	6.09
MC_FBC	**94.57**	**6.09**	4.78	**5.44**

FL denotes Focal Loss, and CE denotes Cross Entropy Loss.

fusion, the learned features from VGG-16 model suffer from a high ACER of over 14%. By contrast, bilinear pooling based methods with richer information have dramatically improved the detection performance - i.e., improving the classification accuracy by over six percentage points and reducing the three error rates (APCER/BPCER/ACER) by more than half on average.

The FBC method achieved better results than the traditional BP and CBP methods due to its more compact and discriminative representation based on sparse coding. More compact representations help to overcome the redundancy and bursti-ness issues of traditional BP. We also present the comparison results of the focal loss function in the proposed MC_FBC scheme with the cross-entropy loss used in the original FBC. For both schemes, the focal loss function can improve the detection results in terms of classification accuracy and error rates thanks to its generalization and robustness by giving more attention to hard and misclassified examples. Overall, the proposed MC_FBC scheme has achieved the highest accuracy and lowest error rates (except the BPCER) on the Protocol III when tested on the WFFD database.

TABLE V
COMPARISONS OF MODEL SIZE AND COMPLEXITY. PARAMS: PARAMETER NUMBER; FLOPs: THE NUMBER OF FLOATING POINT OPERATIONS PER SECOND.

Model	VGG-16		ResNet-50	
	Params	FLOPs	Params	FLOPs
w/o FBC	138.36M	15.48G	25.56 M	4.11G
w/ FBC	15.77M	15.78G	27.71M	4.52G
w/ MC_FBC	15.77M	31.13G	27.71M	8.63G

C. Model Complexity

We have compared the proposed model with the original

TABLE VI
COMPARISON RESULTS (%) ON THE WFFD DATABASE

Method	Protocol I				Protocol II				Protocol III			
	EER	APCER	BPCER	ACER	EER	APCER	BPCER	ACER	EER	APCER	BPCER	ACER
Multi-scale LBP [3]	23.50	27.00	28.50	27.75	31.15	36.15	27.69	31.92	28.91	31.74	25.65	28.70
Image quality [71]	35.50	30.50	39.50	35.00	38.85	39.23	43.46	41.35	41.30	36.96	43.26	40.11
Color LBP [61]	31.50	36.00	28.50	32.25	33.85	30.77	39.61	35.19	31.52	33.26	34.13	33.70
Haralick [10]	30.50	27.50	32.50	30.00	32.69	33.08	36.54	34.81	34.78	28.04	40.00	34.02
Recod [72]	17.00	25.50	14.50	20.00	22.69	25.77	30.77	28.27	21.30	23.91	20.22	22.07
ResNet-50 [73]	16.50	21.00	18.50	19.75	17.31	19.23	21.92	20.58	15.87	17.61	20.43	19.02
VGG-16 [67]	14.50	14.50	18.00	16.25	18.08	13.46	15.77	14.61	14.78	13.48	17.17	15.33
CCoLBP [74]	29.50	26.50	26.00	26.25	28.08	24.62	34.23	29.42	28.04	26.52	29.13	27.83
Noise model [75]	31.00	31.00	48.50	39.75	41.54	41.54	41.15	41.35	38.00	38.04	47.83	42.93
Hybrid ResNet [76]	8.50	9.00	13.00	11.00	11.00	11.38	13.31	12.35	10.90	11.21	13.23	12.22
Human-based	-	20.14	11.86	16.00	-	32.97	17.97	25.47	-	27.39	15.31	21.35
FaceBagNet [77]	15.28	19.50	13.00	16.25	20.33	20.77	21.92	21.35	14.60	17.39	12.39	14.89
CDCN [78]	16.50	17.50	18.00	17.75	23.48	26.54	25.38	25.96	19.22	19.13	19.57	19.35
DeepPixBiS [79]	6.50	8.50	4.00	6.25	8.44	11.54	5.38	8.46	7.82	11.52	4.50	8.04
MC_FBC-VGG-16	5.28	4.35	6.96	5.66	7.34	11.53	**4.23**	7.88	5.32	6.09	4.78	5.44
MC_FBC-ResNet-50	**4.82**	**4.00**	**5.50**	**4.75**	**6.55**	**7.30**	6.54	**6.92**	**4.70**	**5.65**	**3.91**	**4.78**

VGG-16/ResNet-50 and FBC models in terms of model size and complexity. The comparison results within the same input face size are shown in Table V. With one more stream to obtain complementary facial color texture descriptions, the proposed MC_FBC model has higher (almost twice) FLOPs for both VGG-16 and ResNet-50 networks than the original FBC. However, it keeps the same model size as the original FBC, which is still much smaller than that of the VGG-16 model without FBC, and only slightly higher than the ResNet-50 model. Considering the significant performance improvement as shown in Table VI, we deem MC_FBC a good comprised solution between cost and performance.

D. Comparison on the proposed database

1) WFFD database: Several face PAD methods were evaluated on the WFFD database, to show how they can work for super realistic 3D presentation attacks. These PAD methods have achieved promising performance in detecting 2D type or 3D mask presentation attacks. Our benchmark set includes multi-scale LBP [3], image quality assessment based [71], color LBP [61], Haralick features [10], Recod method with outstanding performance in a face spoofing detection competition [72], ResNet-50 based [73], [53], VGG-16 based [67], Chromatic Co-Occurrence of LBP (CCoLBP) [74], noise modeling based [75], Hybrid ResNet [76], FaceBagNet [77], CDCN [78], and DeepPixBiS [79]. The experimental results of all benchmark methods were obtained using the publicly available codes. In addition, we have conducted a controlled human-based detection experiment to test the ability of human vision systems in distinguishing wax figure faces from real ones. In our controlled experiment, 20 volunteers (10 men and 10 women, aged between 23 and 55) were asked to determine whether the face is real or not using our self-developed program. The classification error rates were calculated and averaged as the final result of human-based spoofing detection.

Table VI compares the results of different face PAD schemes. For Protocol I, we can see that the existing face PAD methods for 2D or 3D mask attacks suffer from severe performance degradation with high detection error rates on WFFD, ranging from 8.5% to 48.5%. We attribute the poor performance to high diversity and super realistic attacks in the new database. Among them, the most learned features [72], [73], [67], [76], [77], [78], [79] achieved better results than hand-crafted features [3], [71], [61], [10]. Human-based detection has achieved a lower ACER of 16%, but with higher APCER than BPCER, suggesting that more wax figure faces were mistaken for real ones. The proposed MC_FBC scheme achieved the best results with ACER less than 6% for both backbone networks. Similar performance differences can be observed under Protocol II. However, most algorithms achieved higher error rates for this protocol. Such results are reasonable since recording real and wax figure faces in the same scenario with the same camera results in less difference between real and fake faces. Therefore, it is more difficult to detect spoofing attacks in this homogeneous setting.

The overall results under Protocol III with different face PAD methods have large differences, with the error rates ranging from 3.91% to 47.83%. The best ACER was achieved in the proposed ResNet-50 based MC_FBC scheme due to the highly discriminative features, which significantly outperformed other algorithms and human based detection. Based on pixel-wise binary supervision, the DeepPixBiS method [79] also achieved better results, with all error rates lower than 9%. Human-based detection performs worse than machine-based for all three protocols, which implies that real vs. wax detection is nontrivial for the layperson. The image quality-based [71], and noise modeling-based [75] methods, however, performed worse on the WFFD because of the high diversity of image quality in the proposed database.

2) Failure case analysis: Based on the detection results of WFFD, we further analyze the failure cases in order to

achieve a deeper understanding. In Fig. 7, we have shown the failure cases with high probability in both MC_FBC method and human-based detection results, which visually illustrate the challenges of distinguishing between fake faces and real ones even for human observers. From Fig. 7(a), we note that most false detections with high probability in the proposed MC_FBC method trend to have special face poses, which on the contrary may become the cues for human to detect. More interestingly, when compared with the proposed method, human-based detection was more likely to mistake wax figure faces for real ones for both protocols, as shown in Fig. 7(b) (there are more red dots than green dots). This is in sharp contrast with that in machine-based method in Fig. 7(a) (there are nearly the same number of total green and red dots).

3) WFFD-V database: To detect moving wax figure faces in the newly constructed WFFD-V database, we have randomly selected 10 frames from each video, and the resulting scores from the Softmax classifier were averaged to obtain the final score. Table VII shows that the proposed method achieved the best performance with an accuracy of 94.74% and ACER of 5.23% using either VGG-16 or ResNet-50 network. The last four rows of the table demonstrate the effect of multiple color space fusion in improving the detection results. Compared with the results in Table VI, we can observe that the performance gap among different methods in detecting video-based wax figure faces is smaller than detecting photo-based ones, and most methods achieved lower error rates in this dataset. This can be attributed to the lower diversity of the WFFD-V. Overall, the learning features, especially the bilinear pooling based, patch-based [77], and pixel-wise supervision-based [78], [79], performed better (with accuracy over 91% and ACER less than 9%) than hand-crafted features. The Haralick based method [10] obtained the worst performance with an ACER over 20% in distinguishing between wax figure face videos and real face ones.

E. Intra-database testing on existing databases

1) 3D mask spoofing databases: The spoofing detection results on 3DMAD and HKBU-MARs-V1+ databases, are shown in Tables VIII and IX. Besides the previous benchmark set, we include three face PAD methods proposed for 3D mask spoofing detection- namely, two heartbeat signal-based methods using global or local rPPG-spectrum features [80],

TABLE VII
COMPARISON RESULTS (%) ON THE NEW WFFD-V DATABASE.

Method	Accuracy	APCER	BPCER	ACER
Color LBP [61]	84.21	20.69	10.71	15.70
Haralick [10]	78.95	27.59	14.29	20.94
Recod [72]	83.33	8.93	24.14	16.53
ResNet-50 [73]	87.72	10.34	14.29	12.32
VGG-16 [67]	89.25	11.61	9.91	10.76
CcoLBP [74]	80.70	20.69	17.86	19.27
Noise model [75]	80.04	15.18	24.57	19.87
Hy-ResNet [76]	89.47	10.34	10.71	10.53
FaceBagNet [77]	91.89	9.48	6.70	8.09
CDCN [78]	92.54	7.33	7.59	7.46
DeepPixBiS [79]	93.42	**6.47**	6.70	6.58
BP [58]	91.23	6.90	10.71	8.81
FBC [57]	92.98	6.90	7.14	7.02
MC_FBC-VGG-16	**94.74**	6.90	**3.57**	**5.23**
MC_FBC-ResNet-50	**94.74**	6.90	**3.57**	**5.23**

[16], and the deep dynamic texture-based method [12]. Without publicly available codes, we have directly cited the reported results under the same protocol. It can be seen that the proposed MC_FBC method achieved 0% error rate on 3DMAD database, where several methods performed perfectly. Note that this is because 3DMAD is a relatively easy dataset for spoofing detection (with simple and rigid masks as shown in Fig. 3). On a more realistic HKBU-MARs-V1+ database, our method achieved 3.64% ACER using ResNet-50 as the backbone network, slightly higher than the best result from Haralick features (with 3.24%). We can also observe that all methods have achieved higher error rates on HKBU-MARs-V1+ (see Table IX) than on the 3DMAD database (see Table VIII). Such results are reasonable because the mask spoofing samples in the HKBU-MARs-V1+ are closer to real faces and contain more realistic variations. Based on heartbeat signal analysis, the methods using GrPPG [80] and LrPPG features [16] were affected little by the spoofing quality, showing better detection robustness on the two databases.

2) 2D face spoofing database: Tables X and XI show the comparison results on the 2D face spoofing databases (MSU-USSA and Oulu-NPU). In Table X, we also include

Fig. 7. Failure cases with high probability. (a) in the proposed MC_FBC method (with over 80% of the detection results from 20 randomly chosen epochs using two backbone networks); (b) in human based anti-spoofing detection (with over 80% of the 20 volunteers). Note that images with red dots are wax figure faces (but mistaken for real faces), while images with green dots are real faces (but mistaken for wax faces).

TABLE VIII
COMPARISON RESULTS (%) ON 3DMAD DATABASE

Method	APCER	BPCER	ACER
Haralick [10]	**0.00 ± 0.0**	**0.00 ± 0.0**	**0.00 ± 0.0**
Recod [72]	4.70 ± 19.4	**0.00 ± 0.0**	2.35 ± 9.7
ResNet-50 [73]	**0.00 ± 0.0**	**0.00 ± 0.0**	**0.00 ± 0.0**
VGG-16 [67]	14.21 ± 16.7	3.33 ± 8.5	9.59 ± 11.2
GrPPG [80]	-	-	13.3 ± 13.3[*]
LrPPG [16]	-	-	8.57 ± 13.3[*]
Deep dynamic textures [12]	-	-	1.76[*]
BP [58]	**0.00 ± 0.0**	**0.00 ± 0.0**	**0.00 ± 0.0**
FBC [57]	**0.00 ± 0.0**	1.25 ± 3.5	0.61 ± 1.8
MC_FBC-VGG-16	**0.00 ± 0.0**	**0.00 ± 0.0**	**0.00 ± 0.0**
MC_FBC-ResNet-50	**0.00 ± 0.0**	**0.00 ± 0.0**	**0.00 ± 0.0**

[*] Using reported results.

TABLE IX
COMPARISON RESULTS (%) ON HKBU-MARs-V1+ DATABASE

Method	APCER	BPCER	ACER
Color LBP [61]	24.00 ± 36.5	9.71 ± 9.8	16.72 ± 20.2
Haralick [10]	**3.86 ± 7.6**	2.43 ± 3.8	**3.24 ± 6.8**
Recod [72]	17.14 ± 27.2	15.71 ± 27.4	16.53 ± 28.7
ResNet-50 [73]	17.14 ± 29.2	15.71 ± 26.9	16.43 ± 28.1
VGG-16 [67]	16.33 ± 37.8	7.10 ± 11.2	13.66 ± 19.8
CcoLBP [74]	14.28 ± 34.1	12.85 ± 30.0	13.57 ± 31.9
GrPPG [80]	-	-	16.10 ± 20.5[*]
LrPPG [16]	-	-	8.67 ± 8.8[*]
Deep dynamic textures [12]	-	-	13.44[*]
BP [58]	9.09 ± 30.2	3.55 ± 11.8	6.32 ± 15.6
FBC [57]	9.09 ± 30.2	2.64 ± 8.7	5.86 ± 15.3
MC_FBC-VGG-16	8.33 ± 28.87	**0.00 ± 0.0**	4.17 ± 14.4
MC_FBC-ResNet-50	7.27 ± 24.1	**0.00 ± 0.0**	3.64 ± 12.1

[*] Using reported results. Note that the result of 'Deep dynamic textures' was conducted on a subset of HKBU-MARs-V1+.

TABLE X
COMPARISON RESULTS (%) ON MSU-USSA DATABASE

Method	APCER	BPCER	ACER
Color LBP [61]	3.1 ± 0.8	3.0 ± 0.8	3.1 ± 0.8
Image distortion [24]	3.3 ± 0.7	4.3 ± 2.0	3.5 ± 1.0
Haralick [10]	9.1 ± 0.9	8.8 ± 0.9	8.9 ± 0.9
Recod [72]	3.3 ± 0.4	3.4 ± 1.3	3.3 ± 0.7
ResNet-50 [73]	7.6 ± 0.9	8.9 ± 2.6	8.3 ± 1.0
VGG-16 [67]	27.7 ± 5.5	27.8 ± 2.2	27.8 ± 3.5
Patch-CNN [81]	-	-	0.4 ± 0.3[*]
Depth-CNN [81]	-	-	2.2 ± 0.7[*]
Two stream CNN [81]	-	-	**0.2 ± 0.2**[*]
Deep forest [82]	-	-	1.3 ± 0.5[*]
MC_FBC-VGG-16	**1.0 ± 0.5**	2.9 ± 1.0	1.9 ± 0.5
MC_FBC-ResNet-50	1.5 ± 0.5	**1.6 ± 0.9**	1.5 ± 0.4

[*] Using reported results.

7.6% in the four protocols respectively, only slightly lower than using VGG-16 as the backbone network. The CDCN method achieved superior performance to most methods under the first three protocols with limited variations due to its powerful discrimination ability in extracting intrinsic 2D spoofing patterns [78], [86]. Although the performance of our proposed MC_FBC was not the best under Protocol I with unseen environmental conditions, it has shown good robustness (ranking third, slightly lower than the best result of noise model based method [75]) under Protocol IV, which combines the previous three protocols and is the most challenging scenario. Overall, even though the proposed method is designed for 3D face anti-spoofing, it is still highly competitive when compared against state-of-the-art methods for detecting 2D face spoofing attacks.

F. Inter-database testing

To study the generalization property against unseen attacks, we have conducted inter-database evaluation on both the new WFFD and existing 3D mask spoofing databases. We first show how well existing methods can perform in detecting moving wax figure faces using the still wax figure faces as the training set. The face images in the training subset (2760) in WFFD were used for training, and all 285 videos (each with 10 frames) in WFFD-V dataset were used for testing. Table XII shows larger differences in the accuracy in the range of 52.50% to 86.32% and ACER of 13.72% to 51.26% among different detection methods. The last four rows show the effectiveness of bilinear pooling fusion on improving the performance.

The proposed method demonstrates the best generalizability with the lowest ACER of 17.41% and 13.72% using the VGG-16 and ResNet-50 models respectively. We attribute the outstanding performance to the complementary features learned by the proposed method. Similar to the results in Table VII, most learning feature based methods performed better in detecting video-based wax faces than hand-crafted features.

We have further compared the generalizability of the proposed method to unseen 3D mask spoofing attacks. For a fair comparison, we have followed the protocols in [37], [20]: training on 3DMAD uses random 8 subjects, training on HKBU-MARs-V1+ uses 6 subjects, and both testing uses

two new methods on MSU-USSA database, including two-stream CNN in [81] combining Patch-CNN and depth-CNN, and deep forest with multiscale LBP based method [82]. The proposed MC_FBC method has achieved the ACER of 1.9% and 1.5% using two backbone networks (VGG-16 and ResNet-50) respectively, slightly higher than deep forest based method (with ACER of 1.3%). Using not only full face images, but also local patches extracted from the same face, the two-stream CNN method performed the best with 0.2% ACER in distinguishing the fake from real faces.

For the widely used Oulu-NPU database, we have added several benchmark methods, including two leading methods in face spoofing detection competition held in 2017 [72] (GRADIANT and MixedFAXNet), and five recent works (MILHP [83], CDCN [78], DeepPixeBiS [79], TSCNN [84], and SAPLC [85]). The proposed MC_FBC method based on ResNet-50 model achieved ACERs of 5.9%, 3.8%, 4.9%, and

TABLE XI
COMPARISON RESULTS (%) ON OULU-NPU DATABASE (NOTE THAT THIS DATASET IS FOR 2D SPOOFING ONLY).

Method	Protocol I			Protocol II			Protocol III			Protocol IV		
	AP	BP	ACER	AP	BP	ACER	APCER	BPCER	ACER	APCER	BPCER	ACER
Color LBP [61]	5.0	20.8	12.9	22.5	6.7	14.6	14.2 ± 9.2	8.6 ± 5.9	11.4 ± 4.6	29.2 ± 37.5	23.3 ± 13.3	26.3 ± 16.9
GRADIANT [72]	1.3	12.5	6.9	3.1	1.9	2.5	2.6 ± 3.9	5.0 ± 5.3	3.8 ± 2.4	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5.0
MixedFASNet [72]	**0.0**	17.5	8.8	9.7	2.5	6.1	5.3 ± 6.7	7.8 ± 5.5	6.5 ± 4.6	10.0 ± 7.7	35.8 ± 26.7	22.9 ± 15.2
Recod [72]	3.3	13.3	8.3	15.8	4.2	10.0	10.1 ± 13.9	8.9 ± 9.3	9.5 ± 6.7	35.0 ± 37.5	10.0 ± 4.5	22.5 ± 18.2
Noise model [75]	1.2	1.7	1.5	4.2	4.4	4.3	4.0 ± 1.8	3.8 ± 1.2	3.6 ± 1.6	5.1 ± 6.3	$\mathbf{6.1 \pm 5.1}$	$\mathbf{5.6 \pm 5.7}$
MILHP [83]	8.3	0.8	4.6	5.6	5.3	5.4	$\mathbf{1.5 \pm 1.2}$	6.4 ± 6.6	4.0 ± 2.9	15.8 ± 12.8	8.3 ± 15.7	12.0 ± 6.2
CDCN [78]	0.4	1.7	1.0	**1.5**	1.4	**1.5**	2.4 ± 1.3	$\mathbf{2.2 \pm 2.0}$	$\mathbf{2.3 \pm 1.4}$	$\mathbf{4.6 \pm 4.6}$	9.2 ± 8.0	6.9 ± 2.9
DeepPixBiS [79]	0.8	**0.0**	**0.4**	11.4	**0.6**	6.0	11.7 ± 19.6	10.6 ± 14.1	11.1 ± 9.4	36.7 ± 29.7	13.3 ± 16.8	25.0 ± 12.7
TSCNN [84]	5.1	6.7	5.9	7.6	2.2	4.9	3.9 ± 2.8	7.3 ± 1.1	5.6 ± 1.6	11.3 ± 3.9	9.7 ± 4.8	9.8 ± 4.2
SAPLC [85]	**0.0**	0.8	0.4	2.8	2.2	2.5	4.7 ± 4.2	3.1 ± 3.5	3.9 ± 2.1	11.9 ± 6.9	6.7 ± 5.5	9.3 ± 4.4
MC_FBC-VGG-16	5.7	8.7	7.2	5.1	3.3	4.2	2.9 ± 3.8	9.9 ± 8.9	6.5 ± 4.7	6.3 ± 5.8	10.5 ± 10.9	8.8 ± 5.9
MC_FBC-ResNet-50	3.5	8.3	5.9	3.3	4.2	3.8	3.4 ± 3.5	6.3 ± 4.2	4.9 ± 3.7	5.8 ± 4.5	9.4 ± 9.3	7.6 ± 5.4

All using reported results. 'AP' denotes 'APCER', and 'BP' denotes 'BPCER'.

TABLE XII
INTER-DATABASE TESTING RESULTS (%) ON WFFD-V DATABASE.

Method	Accuracy	APCER	BPCER	ACER
Color LBP [61]	52.50	51.90	42.95	47.42
Haralick [10]	54.04	51.03	40.71	45.87
Recod [72]	67.72	18.62	46.43	32.52
ResNet-50 [73]	60.18	36.29	43.48	39.89
VGG-16 [67]	71.23	38.62	18.57	28.60
CcoLBP [74]	48.77	48.96	53.57	51.26
Noise model [75]	63.38	49.66	23.12	36.39
Hy-ResNet [76]	63.16	53.79	19.29	36.54
FaceBagNet [77]	67.81	40.17	23.93	32.05
CDCN [78]	79.91	20.09	20.09	20.09
DeepPixBiS [79]	80.83	23.10	15.09	19.10
BP [58]	77.50	28.79	15.98	22.39
FBC [57]	78.86	22.59	19.64	21.11
MC_FBC-VGG-16	82.46	24.83	**10.00**	17.41
MC_FBC-ResNet-50	**86.32**	**11.72**	15.71	**13.72**

all subjects. Besides the previous benchmark methods, the latest 3D mask PAD method based on temporal similarity of rPPG (TSrPPG) [20] is added to the benchmark set. Results in Table XIII have justified the robustness of the MC_FBC method, with the second lowest ACER using ResNet-50 as the backbone under both inter-database test settings, while the VGG-16 based MC_FBC performed slightly worse than ResNet-50 model. Most methods have achieved higher error rates using 3DMAD as the training set. The underlying reason is that this database contains less variation in the collected data than the HKBU-MARs-V1+ database. Therefore, the models optimized for this database are not able to generalize well in new acquisition conditions. Due to the good generalizability of liveness cues, the TSrPPG method [20] has achieved the best results using heartbeat signal in the time domain.

G. Visualization Analysis

To show the effectiveness of the proposed MC_FBC on extracting highly discriminative features, we present the visual-ization results on WFFD samples using Grad-CAM [87]. Figure 8 presents the CNN activation heatmaps and corresponding guided Grad-CAMs, which locate the class-discriminative regions and highlight the high-resolution details respectively. Note that Grad-CAM is not suitable for bilinear pooling [88]; thus we use Grad-CAM in the last convolution layer of VGG-16 for all models.

From the first four columns in Figure 8, we can first observe the complementary properties of RGB and YCbCr color spaces in the proposed MC_FBC method. The first four rows show little difference in the attention regions of two color channels for real faces, both focusing on the nose and philtrum area. The last four rows, however, present larger attention differences in wax figure faces. We can observe from the first two rows that RGB color space generally focuses on the eye and upper cheek regions, while YCbCr color mainly focuses on the upper cheek regions. For more confusing wax samples (see the example in the seventh row with poor lighting conditions), YCbCr channel seems to be more robust due to the separation of luminance and chrominance components. However, the RGB channel performs better on wax faces with facial occlusions like eyeglasses (see the example in the last row). This can be attributed to the larger attention regions of RGB channels. Therefore, the fusion of features from complementary color channels can contribute to highly robust and discriminative representations in face anti-spoofing.

Combining with the last four columns in Figure 8, we can compare the class-discriminative localization of the proposed model with FBC and original VGG-16 models. It can be observed from both Grad-CAMs and guided Grad-CAMs that the high activation regions of FBC and MC_FBC models are consistent over real faces, mainly around the nose region. Larger nose regions (with both the apex and bridge of nose) attract the attention of MC_FBC than the FBC model (with just the apex of nose). By contrast, the VGG-16 model focuses on nose or mouth regions over real faces, while for more confusing input faces (see the example with smooth skin in the fourth row), the attention distribution scatters almost randomly, leading to erroneous detection. For the wax face samples in the

TABLE XIII
INTER-DATABASE TESTING RESULTS (%) ON 3DMAD AND HKBU-MARs-V1+ DATABASES.

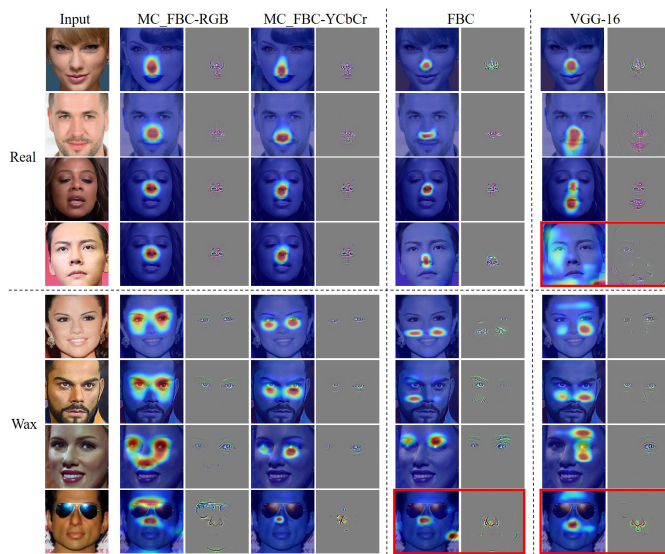| Method	3DMAD → HKBU-MARs-V1+			HKBU-MARs-V1+ → 3DMAD		
	APCER	BPCER	ACER	APCER	BPCER	ACER
Multi-scale LBP [3]	45.00 ± 2.9	43.93 ± 2.9	44.46 ± 2.9	28.76 ± 5.9	28.62 ± 5.8	28.69 ± 5.9
Color LBP [61]	40.00 ± 2.0	39.29 ± 2.5	39.64 ± 2.2	34.70 ± 4.5	34.26 ± 4.8	34.48 ± 4.7
Haralick [10]	30.62 ± 5.9	29.64 ± 5.7	30.13 ± 5.8	21.47 ± 3.8	21.18 ± 3.6	21.32 ± 3.7
ResNet-50 [73]	36.00 ± 8.4	35.14 ± 8.0	35.57 ± 8.2	23.35 ± 8.9	23.06 ± 9.2	23.21 ± 9.1
Recod [72]	32.50 ± 4.6	31.07 ± 3.9	31.78 ± 4.2	26.35 ± 1.4	22.09 ± 5.5	24.22 ± 2.7
VGG-16 [67]	32.50 ± 5.0	31.78 ± 4.5	32.14 ± 3.4	52.71 ± 3.7	40.47 ± 2.9	48.09 ± 2.4
GrPPG [80]	-	-	46.70 ± 3.0	-	-	31.50 ± 3.8
LrPPG [16]	-	-	39.20 ± 0.8	-	-	40.40 ± 2.7
TSrPPG [20]	-	-	$\mathbf{23.50 \pm 0.5}$	-	-	$\mathbf{16.10 \pm 1.0}$
BP [58]	48.08 ± 15.2	$\mathbf{19.86 \pm 22.7}$	33.97 ± 5.5	26.54 ± 9.2	36.07 ± 10.5	31.31 ± 3.0
FBC [57]	34.75 ± 9.9	21.19 ± 18.2	27.97 ± 4.6	29.95 ± 8.1	25.00 ± 12.0	27.50 ± 3.7
MC_FBC-VGG-16	26.25 ± 0.2	25.36 ± 9.5	25.80 ± 9.4	22.79 ± 3.9	20.96 ± 2.0	22.90 ± 4.4
MC_FBC-ResNet-50	$\mathbf{25.00 \pm 2.0}$	25.78 ± 3.8	25.39 ± 2.7	$\mathbf{18.57 \pm 3.6}$	$\mathbf{16.10 \pm 2.6}$	17.34 ± 3.0

3D denotes '3D face reconstruction'.

Fig. 8. Grad-CAM ([87]) attention visualization of the last VGG-16 feature maps corresponding to real faces (the first four rows), and wax figure faces (the last four rows) of the test data samples in WFFD. The first column represents the input faces, and the even columns show the Grad-CAM heatmaps while other odd columns show the guided Grad-CAMs. Red-colored regions of the heatmaps represent highly focused regions, whereas blue regions represent low priority ones. Note that red boxes present error detection.

last four rows, the proposed MC_FBC model mainly focuses on eye areas and upper cheek region, the FBC model focuses on either the upper cheek or the eyebrow regions, while the focus of VGG-16 model seems to be mainly on the upper cheek regions. For the more confusing input face (for example in the last row), both VGG-16 and FBC lead to erroneous detection, while the MC_FBC model is more robust thanks to the learned complementary skin-inspired features.

VI. CONCLUSIONS

To detect realistic 3D face presentation attacks, we have proposed to generate discriminative representations in a fine-grained manner and combine complementary information in multiple color spaces by bilinear coding in this paper. The proposed MC_FBC approach fuses complementary features from two color spaces (RGB vs. YCbCr) extracted via CNN models (VGG-16 and ResNet-50) using factorized bilinear coding. We have also constructed a new database (WFFD) with wax figure faces containing both images and videos with high diversity and large subject size as super realistic face presentation attacks. Extensive experimental results have demonstrated the superior performance of the proposed method in detecting real faces from wax figure faces with not only several existing PAD methods but also human-based spoofing detection. Our method has achieved competitive performance on other 3D mask and 2D face spoofing databases in both intra-database and inter-database testing scenarios. Both the code and databases will be made publicly available at https://github.com/shanface33/Wax_Figure_Face_DB to facilitate the improvement and evaluation of different PAD algorithms.

It should be noted that the best performance under inter-database testing achieved by the proposed scheme still has the error rate of over 10%. How to improve the detection performance deserves further investigation. Super realistic face spoofing attacks are indeed difficult to distinguish from real ones even for humans. We envision that learning-based methods, when combined with liveness cues, are a promising direction to provide effective and generalized spoofing detection in the future. However, as AI technology keeps advancing at a fast pace, it is likely that more challenging spoofing attacks such as Deepfakes will become more powerful. As many people believe, the arm race between spoofing and anti-spoofing will never end.

REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4–20, 2004.

[2] I. O. for Standardization, "Information technology biometric presentation attack detection - part 1: Framework," *ISO/IEC JTC 1/SC 37, ISO/IEC 30107-1:2016(E)*, 2016.

[3] N. Erdogmus and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 1–6.

[4] Y. Xu, T. Price, J.-M. Frahm, and F. Monrose, "Virtual u: Defeating face liveness detection by building virtual models from your public photos," in *25th {USENIX} Security Symposium*, 2016, pp. 497–512.

[5] G. Guo, L. Wen, and S. Yan, "Face authentication with makeup changes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 814–825, 2013.

[6] C. Chen, A. Dantcheva, T. Swearingen, and A. Ross, "Spoofing faces using makeup: An investigative study," in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*. IEEE, 2017, pp. 1–8.

[7] S. Jia, G. Guo, and Z. Xu, "A survey on 3d mask presentation attack detection and countermeasures," *Pattern Recognition*, vol. 98, p. 107032, 2020.

[8] N. Kose and J.-L. Dugelay, "Reflectance analysis based countermeasure technique to detect face mask attacks," in *2013 18th International Conference on Digital Signal Processing (DSP)*. IEEE, 2013, pp. 1–6.

[9] J. Liu and A. Kumar, "Detecting presentation attacks from 3d face masks under multispectral imaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 47–52.

[10] A. Agarwal, R. Singh, and M. Vatsa, "Face anti-spoofing using haralick features," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*. IEEE, 2016, pp. 1–6.

[11] Y. Wang, S. Chen, W. Li, D. Huang, and Y. Wang, "Face anti-spoofing to 3d masks by combining texture and geometry features," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 399–408.

[12] R. Shao, X. Lan, and P. C. Yuen, "Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 923–938, 2018.

[13] Y. Qin, C. Zhao, X. Zhu, Z. Wang, Z. Yu, T. Fu, F. Zhou, J. Shi, and Z. Lei, "Learning meta model for zero-and few-shot face anti-spoofing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 916–11 923.

[14] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3d mask face anti-spoofing with remote photoplethysmography," in *European Conference on Computer Vision*. Springer, 2016, pp. 85–100.

[15] B. Lin, X. Li, Z. Yu, and G. Zhao, "Face liveness detection by rppg features and contextual patch-based cnn," in *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*, 2019, pp. 61–68.

[16] S. Liu, B. Yang, P. C. Yuen, and G. Zhao, "A 3d mask face anti-spoofing database with real world variations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 100–106.

[17] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1713–1723, 2017.

[18] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, 2019.

[19] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 81–89.

[20] S. Liu, X. Lan, and P. Yuen, "Temporal similarity analysis of remote photoplethysmography for fast 3d mask face presentation attack detection," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2608–2616.

[21] R. Shao, X. Lan, and P. C. Yuen, "Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing," in *Biometrics (IJCB), 2017 IEEE International Joint Conference on*. IEEE, 2017, pp. 748–755.

[22] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1713–1723, 2017.

[23] S. Jia, C. Hu, G. Guo, and Z. Xu, "A database for face presentation attack using wax figure faces," *arXiv preprint arXiv:1906.11900*, 2019.

[24] 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.

[25] X. Zhu, S. Li, X. Zhang, H. Li, and A. C. Kot, "Detection of spoofing medium contours for face anti-spoofing," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[26] Y. Kim, J. Na, S. Yoon, and J. Yi, "Masked fake face detection using radiance measurements," *JOSA A*, vol. 26, no. 4, pp. 760–766, 2009.

[27] Y. Wang, X. Hao, Y. Hou, and C. Guo, "A new multispectral method for face liveness detection," in *Pattern Recognition (ACPR), 2013 2nd IAPR Asian Conference on*. IEEE, 2013, pp. 922–926.

[28] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral swir imaging," in *Biometrics (ICB), 2016 International Conference on*. IEEE, 2016, pp. 1–8.

[29] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3d masks," *IEEE transactions on information forensics and security*, vol. 9, no. 7, pp. 1084–1097, 2014.

[30] S. Naveen, R. S. Fathima, and R. Moni, "Face recognition and authentication using lbp and bsif mask detection and elimination," in *Communication Systems and Networks, International Conference on*. IEEE, 2016, pp. 99–102.

[31] N. Kose and J.-L. Dugelay, "On the vulnerability of face recognition systems to spoofing mask attacks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2357–2361.

[32] Y. Tang and L. Chen, "3d facial geometric attributes based anti-spoofing approach against mask attacks," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 589–595.

[33] B. Hamdan and K. Mokhtar, "The detection of spoofing by 3d mask in a 2d identity recognition system," *Egyptian Informatics Journal*, 2017.

[34] S. Bhattacharjee and S. Marcel, "What you can't see can help you–extended-range imaging for 3d-mask presentation attack detection," in *Proceedings of the 16th International Conference on Biometrics Special Interest Group*, no. 231840. Gesellschaft fuer Informatik eV, 2017.

[35] N. Alsufyani, A. Ali, S. Hoque, and F. Deravi, "Biometric presentation attack detection using gaze alignment," in *Identity, Security, and Behavior Analysis (ISBA), 2018 IEEE 4th International Conference on*. IEEE, 2018, pp. 1–8.

[36] A. Ali, S. Hoque, and F. Deravi, "Gaze stability for liveness detection," *Pattern Analysis and Applications*, vol. 21, no. 2, pp. 437–449, 2018.

[37] S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 558–573, 2018.

[38] J. Hernandez-Ortega, J. Fierrez, A. Morales, and P. Tome, "Time analysis of pulse-based face anti-spoofing in visible and nir," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 544–552.

[39] X. Li, J. Komulainen, G. Zhao, P. C. Yuen, and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *International Conference on Pattern Recognition*, 2016, pp. 4244–4249.

[40] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcao, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, 2015.

[41] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 451–460, 2016.

[42] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *International Conference Image Analysis and Recognition*. Springer, 2017, pp. 27–34.

[43] J. Hernandez-Ortega, J. Fierrez, A. Morales, and P. Tome, "Time analysis of pulse-based face anti-spoofing in visible and nir," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 544–552.

[44] A. Ali, N. Alsufyani, S. Hoque, and F. Deravi, "Biometric counter-spoofing for mobile devices using gaze information," in *International*

Conference on Pattern Recognition and Machine Intelligence. Springer, 2017, pp. 11–18.

[45] J. Galbally and R. Satta, "Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models," *IET Biometrics*, vol. 5, no. 2, pp. 83–91, 2016.

[46] J. Xiao, Y. Tang, J. Guo, Y. Yang, X. Zhu, Z. Lei, and S. Z. Li, "3dma: A multi-modality 3d mask face anti-spoofing database," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).* IEEE, 2019, pp. 1–8.

[47] 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.

[48] R. H. Vareto, A. M. Sandanha, and W. R. Schwartz, "The swax benchmark: Attacking biometric systems with wax figures," *arXiv preprint arXiv:1910.09642*, 2019.

[49] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[50] H. Cevikalp, H. S. Yavuz, and B. Triggs, "Face recognition based on videos by using convex hulls," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[51] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 365–374.

[52] X. He and Y. Peng, "Fine-grained visual-textual representation learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 520–531, 2019.

[53] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention lstm networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[54] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[55] J. Shi, I. Alikhani, X. Li, Z. Yu, T. Seppänen, and G. Zhao, "Atrial fibrillation detection from face videos by fusing subtle variations," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[56] S. L. Phung, A. Bouzerdoum, and D. Chai, "A novel skin color model in ycbcr color space and its application to human face detection," in *Proceedings. International Conference on Image Processing*, vol. 1. IEEE, 2002, pp. I–I.

[57] Z. Gao, Y. Wu, X. Zhang, J. Dai, Y. Jia, and M. Harandi, "Revisiting bilinear pooling: A coding perspective," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[58] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[59] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.

[60] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1821–1830.

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

[62] N. Brancati, M. Frucci, G. De Pietro, and L. Gallo, "Dynamic clustering for skin detection in ycbcr colour space," 2016.

[63] Z. Boulkenafet, J. Komulainen, and A. Hadid, "On the generalization of color texture-based face anti-spoofing," *Image and Vision Computing*, vol. 77, pp. 1–9, 2018.

[64] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[65] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 2009, pp. 248–255.

[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[69] 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 International Conference on Automatic Face & Gesture Recognition (FG 2017).* IEEE, 2017, pp. 612–618.

[70] I. O. for Standardization, "Information technology biometric presentation attack detection – part 3: Testing and reporting," *ISO/IEC JTC 1/SC 37, ISO/IEC 30107-3:2017*, 2017.

[71] J. Galbally and S. Marcel, "Face anti-spoofing based on general image quality assessment," in *Pattern Recognition (ICPR), 2014 22nd International Conference on.* IEEE, 2014, pp. 1173–1178.

[72] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, D. Samai, S. E. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, L. Qin *et al.*, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *2017 IEEE International Joint Conference on Biometrics (IJCB).* IEEE, 2017, pp. 688–696.

[73] X. Tu and Y. Fang, "Ultra-deep neural network for face anti-spoofing," in *International Conference on Neural Information Processing.* Springer, 2017, pp. 686–695.

[74] F. Peng, L. Qin, and M. Long, "Ccolbp: Chromatic co-occurrence of local binary pattern for face presentation attack detection," in *2018 27th International Conference on Computer Communication and Networks (ICCCN).* IEEE, 2018, pp. 1–9.

[75] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 290–306.

[76] U. Muhammad and A. Hadid, "Face anti-spoofing using hybrid residual learning framework," in *2019 International Conference on Biometrics (ICB)*, 2019.

[77] T. Shen, Y. Huang, and Z. Tong, "Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[78] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[79] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," *arXiv preprint arXiv:1907.04047*, 2019.

[80] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *2016 23rd International Conference on Pattern Recognition (ICPR).* IEEE, 2016, pp. 4244–4249.

[81] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *2017 IEEE International Joint Conference on Biometrics (IJCB).* IEEE, 2017, pp. 319–328.

[82] R. Cai and C. Chen, "Learning deep forest with multi-scale local binary pattern features for face anti-spoofing," *arXiv preprint arXiv:1910.03850*, 2019.

[83] C. Lin, Z. Liao, P. Zhou, J. Hu, and B. Ni, "Live face verification with multiple instantialized local homographic parameterization." in *IJCAI*, 2018, pp. 814–820.

[84] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 578–593, 2019.

[85] W. Sun, Y. Song, C. Chen, J. Huang, and A. C. Kot, "Face spoofing detection based on local ternary label supervision in fully convolutional networks," *IEEE Transactions on Information Forensics and Security*, 2020.

[86] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 650–651.

[87] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[88] Z. Tang, E. Tian, Y. Wang, L. Wang, and T. Yang, "Non-destructive defect detection in castings by using spatial attention bilinear convolutional neural network," *IEEE Transactions on Industrial Informatics*, 2020.