Polarized Image Translation From Nonpolarized Cameras for Multimodal Face Anti-Spoofing

Yu Tian, Yalin Huang, Kunbo Zhang, *Member, IEEE*, Yue Liu, *Member, IEEE*, and Zhenan Sun, *Senior Member, IEEE*

***Abstract*— In face antispoofing, it is desirable to have multimodal images to demonstrate liveness cues from various perspectives. However, in most face recognition scenarios, only a single modality, namely visible lighting (VIS) facial images is available. This paper first investigates the possibility of generating polarized (Polar) images from VIS cameras without changing the existing recognition devices to improve the accuracy and robustness of Presentation Attack Detection (PAD) in face biometrics. A novel multimodal face antispoofing framework is proposed based on the machine-learning relationship between VIS and Polar images of genuine faces. Specifically, a dual-modal central differential convolutional network (CDCN) is developed to capture the inherent spoofing features between the VIS and the generated Polar modalities. Quantitative and qualitative experimental results show that our proposed framework not only generates realistic Polar face images but also improves the state-of-the-art face anti-spoofing results on the VIS modal database (i.e. CASIA-SURF). Moreover, a polar face database, CASIA-Polar, has been constructed and will be shared with the public at http://biometrics.idealtest.org to inspire future applications within the biometric anti-spoofing field.**

***Index Terms*— Face antispoofing, image translation, polarization, multimodal.**

Manuscript received 22 December 2022; revised 19 May 2023; accepted 17 July 2023. Date of publication 30 August 2023; date of current version 20 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62071468, Grant 62276263, and Grant 62006225; and in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA27040202. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fernando Alonso-Fernandez. *(Corresponding authors: Yue Liu; Kunbo Zhang.)*

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

Yu Tian is with the School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China, and also with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yu.tian@ia.ac.cn).

Yalin Huang is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yalin.huang@cripac.ia.ac.cn).

Kunbo Zhang and Zhenan Sun are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Center for Research on Intelligent Perception and Computing, School of Artificial Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: kunbo.zhang@ia.ac.cn; znsun@nlpr.ia.ac.cn).

Yue Liu is with the School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China (e-mail: liuyue@bit.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2023.3310348, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2023.3310348

I. Introduction

With its noncontact, low cost, and convenient features, face recognition is widely used in daily life in scenarios such as mobile phone unlocking, face swipe payment, and access control. However, existing face recognition systems are susceptible to various spoofing attacks, such as printing attacks, video playback, and 2D/3D masks, leading to challenges in system security. Therefore, improvement of the security of face recognition systems to safeguard user privacy and enhance data security is a pressing issue.

Many face antispoofing (FAS) methods have been proposed, with most approaches focusing on visible light (VIS) images. Early methods have relied on handcrafted features to distinguish between genuine faces and presentation attacks according to texture [1], color [2], and image quality [3], motion cues [4], vital signs [5], and other characteristics. These approaches perform well in controlled environments; however, their performance declines when the environment changes or new attacks emerge. With the development of deep learning, several convolutional neural network (CNN)-based FAS methods have been proposed [6], [7], [8]. These CNN-based methods significantly improve FAS performance by extracting high-level semantic features with deep neural networks. However, despite the usefulness of VIS cues in face liveness detection studies, intrinsic and robust FAS features are difficult to characterize by relying solely on the intensity and RGB information in VIS images. As a result, VIS-based methods typically show poor generalizability when facing several attacks and when image acquisition devices and illumination conditions change.

Some recent work [9], [10], [11], [12] has explored the complementary strengths of different modalities using NIR reflections and the depth of the face structure as supervision to improve FAS performance. First, because genuine faces and spoofing attacks react differently to changes in illumination, reflectance differences are a reliable cue [9]. Second, structural facial depth differences between genuine faces and spoofing attacks are important cues for identifying 2D spoofing attacks [10]. Although these new modal data improve modeling capabilities by analyzing both 3D structures and reflectance cues, depth labels can easily be deceived by increasingly realistic 3D masks, while the infrared approach requires an infrared imaging device with an infrared light source to capture the infrared image. This not only increases the cost but also makes the imaging process prone to overexposure and thus the loss of detailed information in the image.

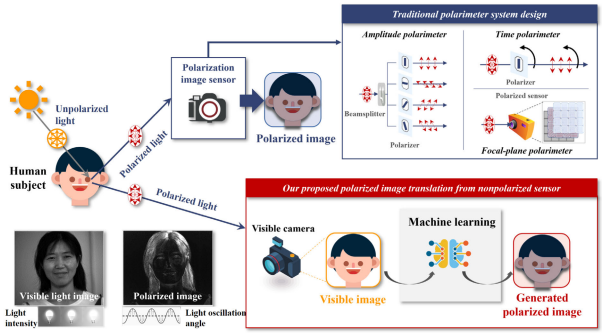


Fig. 1. Outline of the proposed technique. When unpolarized light is reflected from a surface, it becomes partially polarized and the Polar information is usually recorded by Polar imaging systems such as the division of amplitude polarimeters, division of time polarimeters, and division of focal-plane polarimetry. These systems require optical path or chip modifications to existing VIS imaging equipment, adding cost and equipment complexity. We expect to use machine learning to generate Polar images from VIS cameras to establish a mapping between VIS and Polar images to improve the accuracy and robustness of PAD in facial biometrics without changing the existing face recognition devices.

Presentation attacks are always carried out by physical carriers (e.g., paper, LCDs, silicone, rubber) that have different material properties than human facial skin. Upon exposure to natural light, significant differences in the polarization states of the light reflected from various material surfaces are observed, as shown in Fig. 1. Intuitively, sunlight and typical light are both nonpolarized (the vibrational component of the electric field remains constant in all directions). Reflected and transmitted light becomes partially polarized as the light propagates through different material surfaces due to physical properties such as the surface material and roughness [13]. Thus, polarization (Polar), as a passive method, is closely related to the intrinsic physical properties of the material and can provide more robust features for FAS (material classification).

Motivated by the above discussion, several Polar-based FAS methods have been proposed [14], [15], [16]. These works demonstrate that Polar is highly correlated with face material and can uncover the intrinsic differences between genuine and fake faces. Additionally, they verified that face images in the Polar mode are indeed more discriminative than faces in the NIR, depth, and VIS modes. Furthermore, Polar-based FAS methods inevitably require modal consistency across the training and testing phases, similar to NIR-based and depth-based methods. This limits the use of such solutions because the Polar-based approach is not directly compatible with current VIS face anti-spoofing systems. Fig. 1 depicts many typical polarization imaging systems. These designs commonly require the VIS imaging equipment to have its optical paths or chips modified, which incurs a highly significant additional cost in large-scale manufacturing.

In this paper, we focus on two interesting and important questions related to the FAS task: **1) How can intrinsic, robust features be applied to distinguish between genuine and spoofing faces? 2) How can translation between VIS-Polar modes be accomplished such that the existing VIS equipment may also benefit from the Polar mode for better PAD?**

With the rapid development of deep generative models, "recognition by generation" has become a popular research topic in the field of computer vision. For example, the depth maps generated from VIS images in [17] were used for feature learning in a genuine and deceptive face classification task using texture differences between genuine and fake faces in the VIS images and structural differences in the depth maps. References [18] and [19] fused the VIS modalities with the generated NIR modalities to achieve FAS. These methods show that image-to-image translation can to some extent be applied to generate features for new modalities.

Inspired by this, we propose the polarized image translation generator (PTG-Face) detection framework, a new approach for detecting FAS that performs cross-modal translation between VIS and Polar images. The proposed modal translation network is shown in Fig. 3. PTG-Face is divided into two stages: (1) Pairs of VIS and Polar modal images are fed into FC-Net that learns the mapping relationships and feature differences between the two image modalities and transfers the VIS modality images to Polar modality. (2) A dual-modal face PAD network is constructed to mine the intrinsic and robust features of genuine and fake faces by fusing the VIS modalities with the generated Polar modalities. Specifically, we built FC-Net using CycleGAN as a backbone to translate the samples from the VIS modality to the Polar modality. We analyzed the impact on the generated images due to the frequency domain gap in the generation process and explored methods to improve the quality of the generation by narrowing this gap. We propose a frequency domain consistency loss that directly optimizes the generative model in the frequency domain. This will help to make the generated Polar modality closer to the distribution of the real Polar modality. For PAD, translated Polar modalities were used to learn polarization spoofing cues that arise from the material differences between genuine and fake faces. As shown in Fig. 5, these spoofing cues cannot be easily detected in VIS spectra but show significant differences in Polar images. We take advantage of the CDCN's adeptness at describing fine-grained invariant information to propose a dual-stream CDCN that fuses the VIS modalities with the generated polar modalities to learn the intrinsic features of PAD. Similar to the NIR generation method, we do not use the generated Polar modes separately but rather fuse them with VIS features. This design, on the one hand, makes full use of existing VIS imaging equipment and data, and on the other hand, the multimodal approach provides a richer set of PAD features for the FAS task to some extent, obtaining improved accuracy and robustness.

To the best of our knowledge, the proposed PTG-Face method is the first approach to generate Polar images for FAS detection. The main contributions of this work can be summarized as follows.

1. This is the first work to fully explore the advantages of the Polar modality in a VIS-based FAS system, which is achieved using the proposed novel PTG-Face framework.

2. In the PTG-Face framework, we designed a frequency domain-constrained network (FC-Net) module and proposed a novel frequency domain consistency loss as a complement to the existing spatial losses. The generation of polarization-style face images from VIS face images is achieved and preserves the polarization features of both genuine and spoofed faces.

3. We created a dual-stream central differential convolutional network (CDCN) to learn and fuse discriminative presentation attack detection features in translated polarization modalities and real visible light modalities by utilizing the remarkable ability of CDCN to represent fine-grained features invariantly in different environments.

4. As part of this work, we present CASIA-Polar which to the best of our knowledge is the only publicly available polarized face dataset. We have conducted extensive quantitative and qualitative experiments on publicly available visible light datasets and our CASIA-Polar dataset to demonstrate that the proposed method not only shows reliable polarization generation performance but also achieves state-of-the-art face anti-spoofing performance.

II. Related Work

A. Spectrum-Based Methods

With the decreasing cost of multi-spectral sensors and the increasingly popular usage scenarios, some methods for live face detection based on spectral feature analysis have been proposed.

Initially, handcrafted features such as HOG [20], LBP [21], SIFT [22], and SURF [1] were used to distinguish between genuine and fake faces, and support vector machines (SVMs) were used for binary classification. These methods often rely on human liveness cues, which require significant task-aware prior knowledge for their design. Several CNN-based methods have been proposed with the advances in deep learning. Yang et al. [23] were the first to use CNNs in face antispoofing research. Subsequent work, such as [6], [24], and [25], developed CNN-based FAS methods by extracting feature differences between genuine and fake faces, such as texture details, color distortion, and specular reflections. Unfortunately, most of these methods are specifically designed for 2D attacks and perform poorly against challenging 3D and video replay attacks.

Moreover, VIS-based methods are often not sufficiently robust in handling complex and variable attack types and detection scenarios. Therefore, several near-infrared (NIR) [9], shortwave infrared (SWIR) [26], and multispectral-based methods [27] have been proposed for FAS. George et al. [28] proposed a multichannel CNN-based PAD approach and introduced the WMCA dataset that contains data from different channels such as color, depth, NIR, and thermal imaging. Heusch et al. [29] used a CNN model for face PAD on SWIR images. The experimental results demonstrate that the method performs better on SWIR images than on VIS images. Zhang et al. [30] proposed a multispectral PAD method by analyzing the multispectral properties of human skin and materials other than skin and selecting wavelengths with discriminative properties. This use of complementary information between different spectra effectively improves the robustness of FAS systems. However, even though these methods are more powerful than the previous methods, they are spectrum-dependent and have considerable hardware costs, increasing the difficulty of deployment in widespread applications.

B. Physical-Based Methods

Due to the limitations of the existing spectroscopic approaches, some physical analysis-based methods have been proposed. Many advanced approaches have attempted to improve the generalizability of FAS algorithms by learning cues inherent in genuine and fake faces. For example, the use of impulse signal feedback from remote photoplethysmography (rPPG) to detect genuine faces and spoofing attacks is an effective FAS measure. Liu et al. [31] developed a local rPPG model that detects 3D mask attacks by extracting discriminative local heartbeat signals. However, rPPG signals are easily distorted by background noise and object motion. Yao et al. [32] proposed an rPPG-based PAD method that used multiple regions of interest (ROIs) to cover the whole face and applied larger weights to emphasize the regions with richer rPPG signals. However, the performance of such methods is generally not stable due to illumination effects; thus, some PAD schemes based on the fusion of rPPG with NIR, depth [33], [34] and other information have been proposed to improve the robustness of rPPG methods.

Face structure analysis-based methods can also be applied to detect spoofing attacks. Kim et al. [35] captured light distribution changes in light field data and detected edge and ray difference features to implement FAS based on the characteristics of microlens images and subaperture images. Liu et al. [36] combined light field camera data with a CNN to detect subject depths in light field images to distinguish between spoofing attacks and genuine faces. Similar to the approach of Liu et al. [36], Sun et al. [37] used infrared structured light to analyze the surface material and spatial structure of genuine and fake faces to achieve FAS. While these methods explore differences such as pulse detection and geometric structure between genuine faces and presentation attacks, the development of high-precision 3D masks, video playback, and other types of attacks poses substantial challenges to the robustness of such methods.

Different from existing works, Polar-based methods reflect material, texture, and roughness differences between genuine faces and presentation attacks. These features are not affected by the external environment and they are difficult to imitate. Therefore, Polar-based methods have great potential in FAS applications. Polar techniques were first applied to FAS by [14], who demonstrated the feasibility of applying the Polar modality in FAS tasks by showing the differences in Polar images between genuine faces, LCDs, and paper masks. Aziz et al. [15] quantitatively analyzed the intensity of Polar images of genuine faces and paper masks by performing statistical analyses; however, only a few types of presentation attacks were considered, and the experimental analyses were insufficient. Our previous work [16] is the only reported FAS study that uses CNNs combined with Polar images, demonstrating that the Polar modality has stronger robustness and generalizability than the VIS modality for the same network structure. Moreover, [16] proposed the CASIA-DoLP FAS dataset (the predecessor of the CASIA-Polar dataset) for the Polar modality. Although these Polar-based methods show strong generalizability and robustness in PAD tasks, these methods require the use of specialized Polar imaging

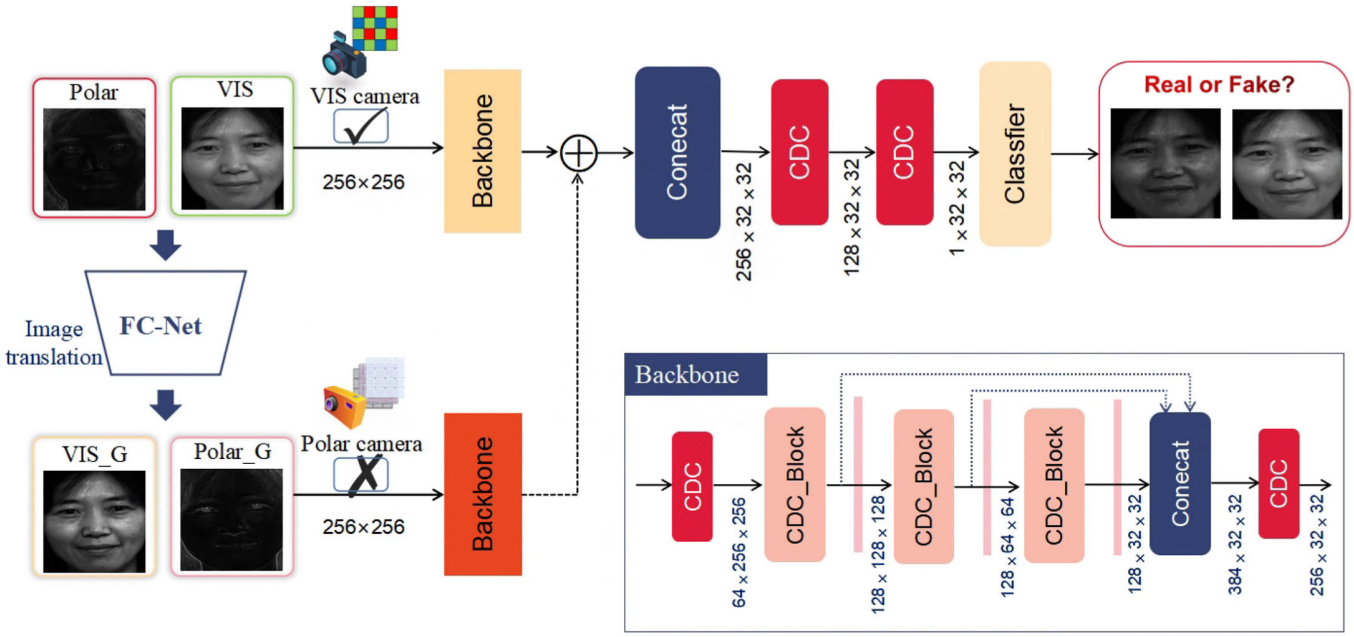


Fig. 2. The overall framework of PTG-Face. This work is a novel face antispoofing method based on existing VIS cameras that translates VIS images to Polar images via FC-Net and performs face antispoofing via a dual-stream CDCN network without utilizing any polarization imaging equipment during the testing process.

equipment to obtain Polar information, which is difficult to apply in existing VIS-based recognition systems.

If we can translate between the VIS and Polar modalities, we can take advantage of the generalizability and robustness of the Polar modality while only using VIS devices.

C. Cross-Modality Image Translation

In recent years, several generative models have shown promise in image translation studies [38], [39], [40].

Image translation is a constrained image generation process that maps an image from one modality to another. This approach is widely used in applications such as face image generation, attribute editing, and superresolution studies. Image translation is achieved due to the powerful ability of generative adversarial networks (GANs) to fit various data distributions. Isola et al. [41] proposed a "pix2pix" framework based on conditional GANs and used the input image to generate the corresponding output image. Zhu et al. [40] extended the pix2pix framework with CycleGANs to learn the mapping function between two unpaired domain images, X and Y, for image generation. On this basis, several generative networks have been proposed. The starGAN [42] algorithm was developed for unpaired multidomain image-image translation. The styleGAN [43] is a style-based generator that automatically separates facial attributes.

GANs have demonstrated great potential in the generation of stylized modalities, such as cross-spectral modalities [44], Sketch-Photo [45], and Profile-Frontal Photo [46]. Moreover, some novel GANs have been considered in FAS applications. [17] generated depth maps from VIS images and detected 2D spoofing attacks by using the face structure information in the depth map. Jiang et al. [19] proposed a FAS approach that fused the VIS and generated NIR modalities. Liu et al. [18] proposed an effective strategy for generating NIR modalities to assist VIS-based FAS systems based on [19], using a partially shared fusion strategy to learn the complementary information of multiple modalities.

Despite their great success, in most cases, a gap between the real and generated images is still observed, particularly in the frequency domain. The frequency domain gap between real and fake images has been attributed to some inherent biases in the neural networks when applied to the generation task. In this paper, we design an FC-Net to learn the VIS-Polar face modal translation that translates face images from VIS to the Polar modality. Additionally, we design frequency domain constraints to narrow the gaps between the generated image and the real image in the frequency domain. This loss is complementary to the existing spatial loss in CycleGAN for adjusting the distribution of the generated Polar modalities to approximate the spatial distribution of the real Polar modalities.

III. Proposed Method

The goal of our approach is to exploit the full potential of the VIS and Polar modalities in order to improve FAS performance. Accordingly, our approach is divided into two parts to address the following issues: (1) how to generate Polar modal information from nonpolarized data and (2) how to use the generated Polar data for FAS. These two parts are presented separately in the following subsections.

A. Overall Architecture of the PTG-Face Framework

In this section, we introduce the PTG-Face framework that focuses on exploiting the Polar modality to improve the performance of VIS-based PADs. Our PTG-Face framework is depicted in Fig. 2. We first trained the generator network with

pairs of heterogeneous data (VIS and Polar). We designed FC-Net with CycleGAN as the backbone to generate Polar images according to the VIS images. Frequency domain gaps between the generated image and the real image often exist due to some inherent bias in neural networks, resulting in distortion of the generated image [47]. In this paper, we improve the quality of image translation by constraining the frequency domain gap between the real and generated modalities. The framework uses sample pairs of VIS and Polar images as input, and the VIS to Polar mode translation is achieved by adversarial training with a discriminator against the generator. This operation generates Polar data and does not require the involvement of any Polar imaging equipment. These generated Polar images and the original VIS images are input into the FAS network. The network, in order to extract more fine-grained and robust features for PAD, consists of a dual-stream central difference convolutional network (CDCN) with independent component channels for the VIS and generated Polar images that achieve live versus prosthetic discrimination by performing feature fusion.

B. Theoretical Foundation of Polarization Approach

In the Polar modality, the differences in Polar characteristics between genuine faces and presentation attacks are apparent. The visualization results in Fig. 5 and previous research [48] both show that the differences between faces and presentation attacks are easier to distinguish in Polar images than in VIS images.

From Maxwell's electromagnetic field theory, it is known that for an arbitrary plane light wave, its light vector can be decomposed into two mutually orthogonal components, the s-wave component vertically to the incident plane and the p-wave component parallel to the incident plane. When an unpolarized beam of light interacts with a surface and is reflected or transmitted, the amount of s- and p-waves will change depending on the surface properties such as surface material, texture, and roughness, causing the unpolarized light to become partially polarized. In other words, the polarization of the light reflected from a surface is determined by the surface material. In FAS studies, diffuse reflection is dominant for most surfaces.

According to the Fresnel formula, the degree of polarization (DoP) at each point $\mathbf{u}$ in the perspective case can be expressed in terms of the refractive index η as follows:

$$\frac{1}{\mathrm{DoP}(\mathbf{u})} = \frac{4\cos\theta(\mathbf{u})\sqrt{\eta^2 - \sin^2\theta(\mathbf{u})}}{\left(\eta - 1/\eta^2\right)\sin^2\theta(\mathbf{u})} + \frac{2\left(1+\eta^2\right)}{\left(\eta - 1/\eta^2\right)\sin^2\theta(\mathbf{u})} - \frac{\left(\eta + 1/\eta^2\right)}{\left(\eta - 1/\eta^2\right)} \quad (1)$$

where θ is the zenith angle to the target surface that varies with the texture and roughness of the target, and $\mathbf{u} = (x, y)$ is a location in the polarization image.

Equation (1) relates the DoP to a function $f(\eta, \theta)$ that depends on the microsurface structure and refractive index of the target. A detailed description of the DoP images can be found in [49]. Moreover, in the supplementary material, we show a more detailed derivation of the DoP images.

As mentioned previously, reflections and refractions on different material surfaces produce idiosyncratic polarized light. While an in-depth quantification of the Polar effects of different materials is beyond the scope of this paper, we note that Polar properties are highly dependent on the material (η) and texture (θ). Although spoofing faces have become increasingly realistic, it is difficult for a spoofing face to exactly match the material and texture of a genuine face. Therefore, spoofing attacks can be distinguished from genuine faces by examining the differences in Polar properties.

However, the acquisition of Polar information requires the use of Polar imaging equipment, which is difficult in existing PAD frameworks that typically utilize only VIS equipment. We hope that generating Polar images based on VIS data acquired by existing VIS equipment, reducing costs (Polar cameras cost approximately 10 times more than VIS cameras with the same imaging parameters) while improving PAD performance.

C. Polarization Translation

In this paper, we generate Polar images instead of acquiring Polar data. Furthermore, we fuse the generated Polar images with real VIS images to improve the PAD performance. In order to transform genuine faces and spoofing attacks from the VIS modality to the Polar modality, we designed the frequency domain-constrained CycleGAN network (FC-Net).

FC-Net uses CycleGAN as a backbone and has a ring-shaped structure with two generators, G and F, and two discriminators, D_X and D_Y. The X-domain image is a VIS-domain image, and the Y-domain image is a Polar-domain image. The X-domain image is passed through generator G to produce the Y-domain image $G(X)$. The X-domain image $F(G(X))$ is then reconstructed by generator F. The Y-domain image is passed through generator F to produce an image in the X domain, $F(Y)$. Similarly, $F(Y)$ is reconstructed through generator G to produce a Y-domain image, $G(F(Y))$. The discriminators D_X and D_Y ensure that the appropriate image style is converted.

As shown in Fig. 3, to generate Polar faces $G(X)$ and ensure that they have the same characteristics as real Y, we add the frequency domain consistency loss $\mathcal{L}_f$ to the constraints imposed by the cycle consistency loss $\mathcal{L}_c$ and adversarial loss $\mathcal{L}_a$. The loss function of our goal can be expressed as:

$$\mathcal{L} = \mathcal{L}_c(G, F) + \mathcal{L}_a(G, D) + \mathcal{L}_f(X_f, Y_f) \quad (2)$$

where the cycle consistency loss $\mathcal{L}_c$ is a regularizer that drives G and F to be consistent with each other in the source modality but not in the target modality. In other words, it aims to minimize the difference between the reconstructed image $G(F(Y))$ $(F(G(X)))$ and the original input image Y (X). The specific expression is shown below.

$$\mathcal{L}_c(G, F) = \mathbb{E}_{X\sim p_{\mathrm{data}}(X)}\left[\|F(G(X)) - X\|_1\right] + \mathbb{E}_{Y\sim p_{\mathrm{data}}(Y)}\left[\|G(F(Y)) - Y\|_1\right] \quad (3)$$

The adversarial loss $\mathcal{L}_a$ is used by discriminator D to distinguish between the original $X(Y)$ domain image and the generated $F(Y)$ $(G(X))$ domain image. The adversarial loss

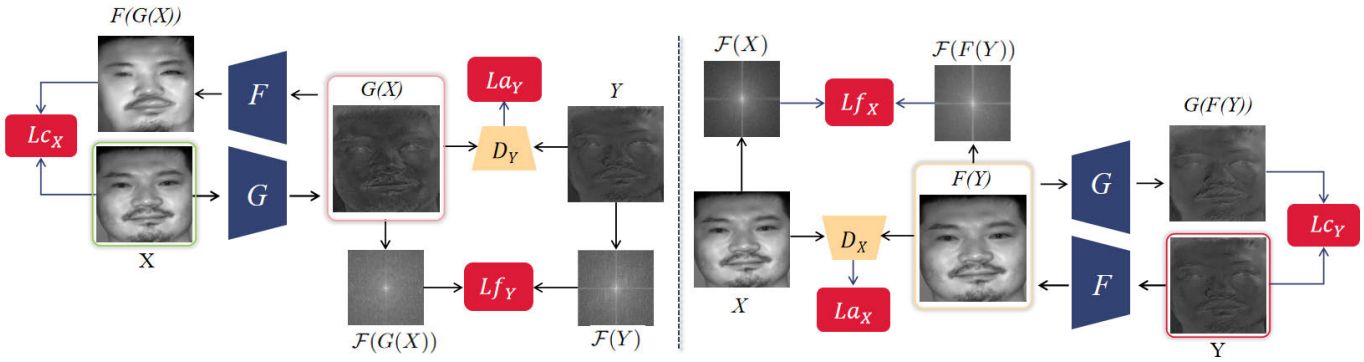


Fig. 3. The overall framework of FC-Net. In the left half of the figure above, the VIS image X is fed into generator G to generate the polarization domain picture $G(X)$, and then the image $G(X)$ is fed into generator F to generate the original domain image $F(G(X))$. The purpose of generating $F(G(X))$ is to use it with the input true image X to calculate the $\mathcal{L}_c$ Loss. The right half is similar, with the input being the polarization domain image Y, which generates the VIS image $F(Y)$ and the polarization image $G(F(Y))$, respectively. During the generation of $X \rightarrow G(X) \rightarrow F(G(X))$ and $Y \rightarrow F(Y) \rightarrow G(F(Y))$, we imposed a $\mathcal{L}_a$ Loss by training the discriminator D_Y (D_X) against the generator G (F). In addition, a frequency domain consistency loss was imposed between the generated image $G(X)$ ($F(Y)$) and the real image X (Y) to ensure the consistent frequency domain distribution.

$\mathcal{L}_a$ directly uses the expression proposed in [50], which is shown below.

$$\begin{aligned}\mathcal{L}_{aX}(F, D_X, X, Y) &= \mathbb{E}_{X\sim p_{\text{data}}}(X)\left[\log D_X(X)\right] \\ &\quad + \mathbb{E}_{Y\sim p_{\text{data}}(Y)}\left[\log\left(1 - D_X(F(Y))\right)\right] \quad (4)\\ \mathcal{L}_{aY}(G, D_Y, X, Y) &= \mathbb{E}_{Y\sim p_{\text{data}}}(Y)\left[\log D_Y(Y)\right] \\ &\quad + \mathbb{E}_{X\sim p_{\text{data}}(X)}\left[\log\left(1 - D_Y(G(X))\right)\right] \quad (5)\end{aligned}$$

We aim to ensure that the generated Polar image $G(X)$ has as consistent a feature distribution as possible with the real Polar image Y.

In this case, the generated Polar image will effectively work as the real Polar image for FAS. However, the generative model has difficulty in maintaining important frequency information as it tends to generate frequencies with higher priority [47]. This will result in a frequency domain gap between the generated image and the real image. Improvement of the quality of image generation through the frequency domain is still largely unexplored.

In this paper, we explore the methods to improve the generated quality by narrowing the gap. We performed the two-dimensional discrete Fourier transform to convert the image into its frequency representation.

$$F(u, v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x, y)\cdot e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \quad (6)$$

where $M \times N$ is the image size, (x, y) denotes the coordinates of the image pixels in the spatial domain, $f(x, y)$ is the original image, (u, v) are represents the coordinate of a spatial frequency on the frequency spectrum, $F(u, v)$ is the result after the Fourier transform, and e and i are the Euler's number and imaginary units, respectively.

It can be found that in Equation (6), $F(u, v)$ depends on the sum of the function of each image pixel in the spatial domain.

We suppress different regions in the spectrum and visualize their physical significance in the spatial domain to simulate the loss of spectrum during generation, as shown in Fig. 4. In Fig. 4, we suppress the low-frequency information of the spectrum (column 2), and the image edges are better preserved, but the overall contrast is reduced. In column 3, the lack of high-frequency information results in a blurred image and typical ringing artifacts. In the fourth and fifth columns, we selectively remove certain frequencies from the image, and it is observed that both images produce varying degrees of distortion, along with the usual checkerboard artifacts.

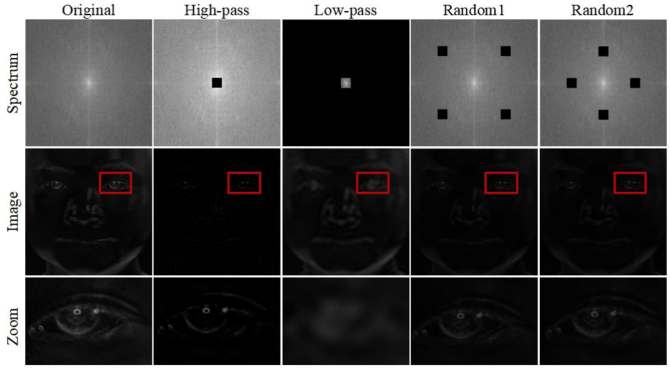


Fig. 4. Standard bandlimiting operations on the frequency spectrum with the origin (low frequencies) center shifted and respective images in the spatial domain. These manual operations can be regarded as a simulation to show the effect of missing frequencies.

Clearly, the losses in different regions of the spectrum will produce different artifacts in the image. Therefore, we deduce that constraining the frequency domain during the generation process can reduce artifacts and improve the quality of the generated image. Therefore, a novel frequency domain consistency loss $mathcal L_f$ is proposed to encourage the mapping of the $G(X)$ (or $F(Y)$) modality to the Y (or X) modality.

The $\mathcal{L}_f$ in FC-Net is the sum of $\mathcal{L}_{fx}$ and $\mathcal{L}_{fy}$, where $\mathcal{L}_{fx}(X, F(Y))$ denotes the difference between X and $F(Y)$ in the frequency domain and $\mathcal{L}_{fy}(Y, G(X))$ denotes the difference between Y and $G(X)$ in the frequency domain. The objective function of our goal can be expressed as:

$$\begin{aligned}\mathcal{L}_{fx}(X, F(Y)) &= \mathbb{E}_{X\sim p_{\text{data}}(X)}\left[\|\mathcal{F}(X) - \mathcal{F}(F(Y))\|_1\right] \\ \mathcal{L}_{fy}(Y, G(X)) &= \mathbb{E}_{Y\sim p_{\text{data}}(Y)}\left[\|\mathcal{F}(Y) - \mathcal{F}(G(X))\|_1\right] \quad (7)\end{aligned}$$

The frequency domain constraint will encourage finer-grained feature consistency between the generated image and the real image, which is better demonstrated in Figs. 6 and 7.

Our source VIS and target Polar images both contain images of genuine faces and spoofing attacks. We perform only modality translation in FC-Net, and by learning, we train FC-Net using paired VIS and Polar images of genuine faces and spoofing attacks as input data. We perform feature fusion on paired VIS and Polar images to take full advantage of both types of images.

D. Face Anti-Spoofing

After the cross-modal face translation, we will obtain the generated Polar modality (Polar_G) from the VIS modality. Pairs of VIS source images and Polar_G images are fed into a dual-stream CDCN network to learn the PAD features.

CNN-based approaches focus on deeper semantic features, are weak in describing detailed intrinsic information between living and spoofing faces, and are prone to fail when dealing with heterogeneous images (e.g., images captured when lighting and camera conditions change). Because spatial differential features are strongly illumination-invariant and contain finer-grained spoof cues, inspired by the traditional LBP difference idea, the CDCN [55] is proposed.

The central difference convolution (CDC) can effectively improve the representation of invariant fine-grained features in different environments. Specifically, sampling and aggregation are the two steps that make up the CDC. The sampling phase resembles vanilla convolution. The CDC tends to aggregate the sampled values' center-oriented gradient during the aggregation stage, giving the CDC a richer representation of detailed features compared to the conventional convolution.

The CDC operator has performed well in the VIS, depth, and NIR modalities [55], [56], but there is no relevant investigation in Polar modalities. We extend the state-of-the-art single modal network CDCN to a dual-modal version for learning PAD features in VIS and Polar_G modes.

We designed our network according to two considerations. First, VIS images are rich in intensity-level semantic information, and Polar_G images retain pixel-level intrinsic feature information. Thus, by using CDCN for feature fusion between VIS and Polar_G images, the information in two modalities can be fully utilized for FAS. Second, the CDCN reliability represents detailed intrinsic patterns and is thus more suitable for Polar feature extraction. The effective learning of these fine-grained "discriminative" and "robust" features is essential for improving PAD performance.

As a result, we use the Polar_G and the original VIS modality X as bimodal inputs to learn the FAS features of the samples. We adopt the configuration CDCN as the backbone network for both modality branches, and the details are shown in Fig. 2. For both modalities, given a face image of size 256×256, the CDCN can extract multilevel (low-level, medium-level, and high-level) fusion features [55]. The backbone networks of the two modality branches are not shared. Therefore, each branch learns modality-aware features independently. Finally, the two head layers aggregate the multimodal features and predict the genuine and false face categories using a classifier.

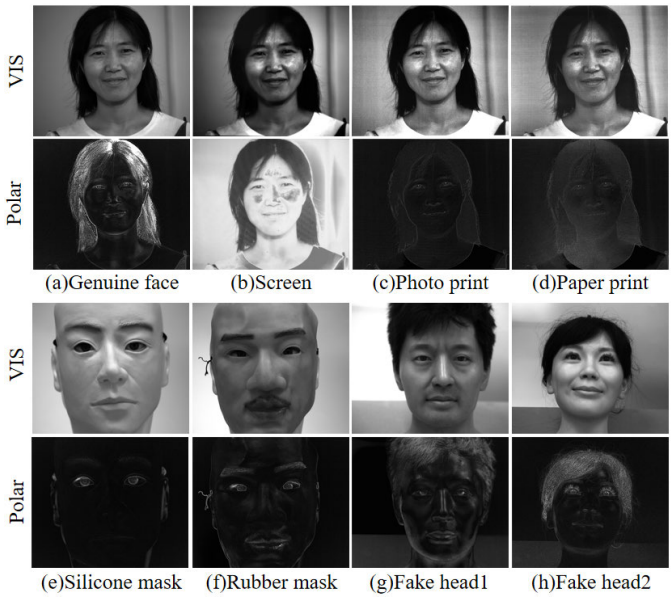


Fig. 5. The attacks present in our extended CASIA-Polar dataset include (a) genuine face, (b) computer screen replay, (c) photo paper prints, (d) A4 paper prints, (e) silicone masks, (f) rubber masks, and (g) and (h) custom-made prosthetic heads using real hair and silicone.

IV. Experiments

In this section, we conduct extensive experiments to demonstrate the efficacy of our approach. The following sections describe the experimental setup, implementation details, and results.

A. Experimental Setup

1) Dataset: While our method requires only VIS data during the testing phase, the model training phase requires Polar data. Unfortunately, although numerous large-scale benchmark datasets have been proposed for FAS research, such as CASIA-SURF [53], CeFA [54], OULU-NPU [51], SiW [52], HiFiMask [57] and WMCA [28], the images and video streams in these datasets are mainly VIS, NIR, and depth modalities. Thus, we cannot easily train our algorithms on existing benchmark datasets, as we need real-time Polar data. To address this need, we collected a Polar FAS dataset known as CASIA-Polar, as shown in Table I. CASIA-Polar is an extension of our CASIA-DOLP. Compared to CASIA-DOLP, the number of subjects was increased to 121 and the amount of attack data was greatly expanded, while the variety of presentation attacks was expanded by customizing lifelike counterfeit heads.

To the best of our knowledge, the CASIA-Polar dataset is the only publicly available Polar face dataset for FAS. We acquired the data using a Lucid Phoenix PHX050S-P polarized camera with the Sony polarization sensor to capture paired VIS images and Polar images. The CASIA-Polar dataset consists of two-dimensional and three-dimensional attack subsets. The 2D attack subset includes photo paper prints,

TABLE I
COMPARISON OF PUBLIC FACE ANTISPOOFING DATASETS

Dataset	Year	#Subject	#Num	Attack	Modality	Device
OULU-NPU [51]	2017	55	5940	Print,Replay	VIS	VIS Camera
SiW [52]	2018	165	4620	Print,Replay	VIS	VIS Camera
CASIA-SURF [53]	2019	1000	21000	Print,Cut	VIS/Depth/IR	Intel RealSense
CeFA [54]	2019	1607	23538	Print, Replay,3D print mask, 3D silica gel mask	VIS/Depth/IR	Intel RealSense
WMCA [28]	2019	72	6716	Print, Replay,2D/3D mask	VIS/Depth/IR/Thermal	RealSense/STC-PRO
CASIA-DOLP [16]	2019	108	10697	Photo/Paper Print,Replay, Silicone Mask,Rubber Mask	VIS/Polar	Polar Camera
CASIA-Polar(Ours)	2022	121	22174	Photo/Paper Print,Replay,Fake Head, Silicone Mask,Rubber Mask	VIS/Polar	Polar Camera

Fig. 6. Comparison with CycleGAN-generated results. The first row shows the VIS modality, the second row shows the results generated using CycleGAN, the third row shows the results generated by our FC-Net, and the last row shows the real acquired polar image.

A4 paper prints, and computer screen replay attacks, and the 3D attack subset includes silicone masks, rubber masks, and custom-made prosthetic heads. A total of 121 subjects were recruited for this study. Three types of 2-D attack samples were collected for each subject. Real samples were collected for each subject at the distances of 1, 2, and 3 m from the collection device. Fig. 5 shows several representative examples of the demonstrated attacks in the CASIA-Polar dataset. Due to the strong correlations between Polar features and the target's intrinsic physical characteristics, such as material and texture traits, the Polar modality exhibits prominent discrimination between genuine and artificial faces.

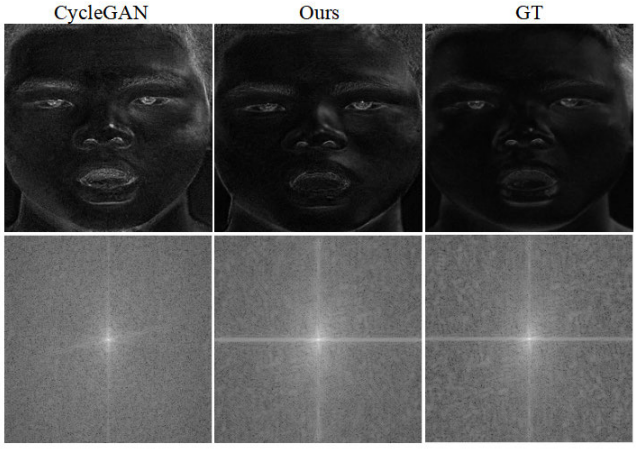
We assessed the CASIA-Polar dataset using four protocols: cross-illumination, cross-attack, cross-distance, and cross-face pose and expression change tests. Because we aimed to improve FAS performance by generating Polar images, all of the experiments were conducted according to protocol 4 (cross-face pose and expression change).

For a quantitative evaluation, we performed a Polar face generation test on the CASIA-SURF dataset and evaluated the FAS performance of PTG-Face on this dataset. For a fair comparison with earlier techniques, we only used VIS modality data from CASIA-SURF; hence, our experiments employed the official test protocol 1 (within-modal evaluation) for this dataset.

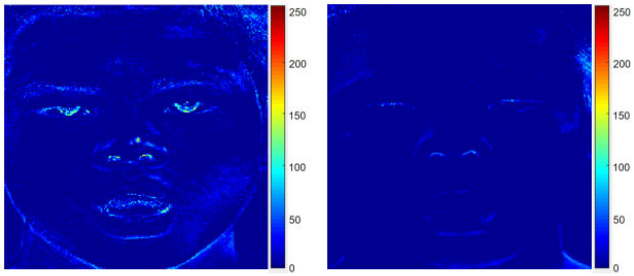
2) Evaluation Metrics: To evaluate the algorithms, we used the ISO/IEC 30107-3 metrics [58], the Attack Presentation Classification Error Rate (APCER), the Bonafide Presentation Classification Error Rate (BPCER), and the Average Classification Error Rate (ACER) metric on the evaluation set.

In addition, the F_score and accuracy rate were compared to assess the classification accuracy of the algorithms.

3) Strategy for Training and Testing: The proposed PTG-Face framework was implemented in parallel on eight NVIDIA GTX-3090 GPUs. The cropped face region was resized to 256×256, and homomorphic filtering was used to remove



(a) Comparison of generated results.



(b) Generation errors of CycleGAN (left) and our approach (right).

Fig. 7. In (a), the first row from left to right, the generated CycleGAN results, the results generated by our method, and the true images captured by the polarization camera. The second row shows the corresponding frequency domain images. For the CycleGAN generation model, there is a larger frequency domain gap between the real image and the generated image, and important frequencies are lost during the generation process, leading to blurry images. Additionally, the loss of high-frequency information leads to serious distortion of image details, as shown in the left panel in (b). By contrast, our FC-Net can maintain maximum frequency domain consistency with the real Polar image due to the addition of $\mathcal{L}_f$, resulting in lower generation errors, as shown in the right panel in (b).

lighting effects. In all experiments, the models were trained for 200 epochs using the Adam solver. All models were trained with a batch size of four and an initial learning rate of 0.0002. We kept the learning rate constant for the first 100 epochs and then linearly decayed the learning rate to zero over the next 100 epochs. After training, the parameters were saved as modal translators for future experiments.

B. Experimental Results

1) Generation of Results Visualization and Analysis: We qualitatively and quantitatively evaluated the Polar image generation results. The collected Polar data were used as the ground truth. To verify the trained FC-Net model's capacity to generate Polar data by the VIS data, a comparison of the results obtained by the CycleGAN [40] and FC-Net models for consistent input data is provided in Fig. 6.

Compared to CycleGAN, FC-Net generates images with higher quality and better preservation of image sharpness and signal-to-noise ratio. The generated Polar images are more accurate and detailed. The results produced using CycleGAN directly have a higher overall contrast, but the generated images have more noise and some local distortion (such as the position of the eyes and mouth).

Both CycleGAN and FC-Net generate near-Polar style images and retain the same face structure information as the source input image. However, FC-Net generates higher-quality images compared to CycleGAN which produces results with higher overall contrast but generates images with more noise and local distortion (e.g., eye and mouth positions). By contrast, FC-Net results are more accurate than the CycleGAN and have richer detailed information, e.g., details in the regions such as eyes and nostrils, as well as the color and texture of the facial skin.

The priority of fitting particular frequencies in a network varies throughout the training, often going from low to high [59]. We found that CycleGAN tends to eschew difficult-to-synthesize frequency components, i.e., hard frequencies, and converge to a lower point during generation, as other generative networks do. As a result, it is difficult for the model to maintain important frequency information, resulting in a large error between the generated effect, and the true value. We provide a quantitative and qualitative evaluation, as shown in Fig. 7. The generated results of CycleGAN retain the structural information of the face but lose some frequency domain information so that the generated images lack sufficient texture details.

To prevent the loss of frequency details by the network, we add $\mathcal{L}_f$ loss to the generation process to supervise the frequency domain information so that the results of the method proposed in this paper agree with the true values, particularly in detailed regions such as nostrils and eyes. Fig. 7(a) depicts a visual comparison of the generated results with the original Polar image. The errors of CycleGAN and our method are shown in Fig. 7(b). It can be easily concluded that the frequency domain agreement between our generated results and the real image is higher, the generation error is lower and therefore the detailed features of the image are better preserved (quantitative analyses of other samples can be found in the supplementary material). These results demonstrate the effectiveness of frequency domain consistency loss that can increase the sensitivity of our generation network to detailed regions.

In addition to genuine faces, we generated images of several types of presentation attacks. The VIS image of the presentation attack was generated as a Polar image (Polar_G) via FC-Net, and the Polar_G images were compared with the Polar images captured by a Polar imaging device. Examples of the captured real Polar image and the Polar_G images are shown in the second and third rows of Fig. 8, respectively. As mentioned above, the generated presentation attacks and real Polar images contain some Polar characteristics of the corresponding class. For example, the skin of a genuine face appears smoother than the presentation attack. Genuine faces have sharper hair details, whereas presentation attacks have blurrier hair details. Furthermore, the generated Polar images and captured Polar images have some differences in visual. For example, the screen replay attack results shown in column 2 of Fig. 8 demonstrate chromaticity differences between the generated and captured Polar images, which may be due to the small

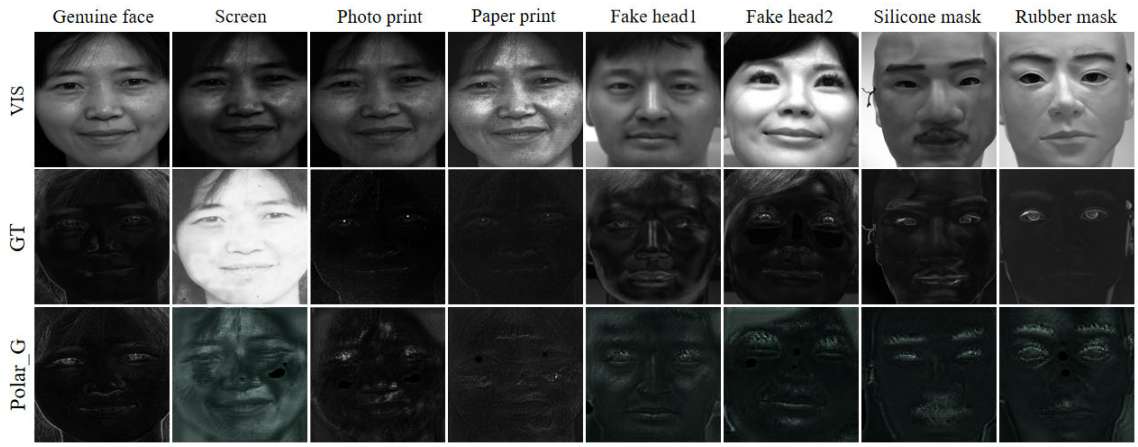


Fig. 8. A visual presentation of the CASIA-Polar generation results, including a real face and seven attack types, e.g., computer screen, photo print, A4 paper print, two dummy heads, a silicone mask, and a rubber mask. The second row shows real polar images captured by the polarization camera, and the third row shows the results translated by our FC-Net.

TABLE II
STATISTICS OF THE CASIA-POLAR DATASET

Modality Setting	Method	Accuracy(%)	F_score(%)	APCER(%)	BPCER(%)	ACER(%)	Params
VIS	ResNet	94.89	95.04	5.97	0.94	3.45	11.171M
Polar	ResNet	99.03	99.43	3.66	0.08	1.87	
Polar_G	ResNet	98.06	95.43	5.64	0.06	2.85	
VIS	CDCN	92.16	93.62	7.44	0.60	4.02	**4.583M**
Polar	CDCN	96.59	97.25	3.47	0.72	2.09	
Polar_G	CDCN	97.64	96.12	4.16	1.42	2.79	
VIS & Polar_G	**Ours**	**99.21**	**99.56**	**0.48**	**0.02**	**0.25**	6.572M

number of data samples in this category. In addition, the colors and textures in some of the generated Polar images are not as smooth and natural as those in the captured Polar images. There are two reasons for this phenomenon. First, there is a large intraclass variation in the spoofing attack itself, such as differences in the display of the same picture by monitors of different resolutions. Second, the number of genuine faces and spoof attacks in CASIA-Polar is approximately 1:1, leading to a low number of each class in the spoofing attack.

2) FAS Results on CASIA-Polar: Since the effectiveness and feasibility of FC-Net-generated Polar images are demonstrated, we compare the performance of the Polar_G images in FAS tasks on the CASIA-Polar dataset. Specifically, we compare the performance of the CASIA-Polar dataset when using VIS images, Polar images, and Polar_G as input data. The PTG-Face model is used for genuine-false face classification to evaluate the use of VIS and Polar_G images as input data.

We studied seven models separately on protocol 4 and report the results in Table II. Comparing ResNet, CDCN, and our PTG-Face method under three modalities, namely, VIS, Polar, and Polar_G, we find that the Polar_G modality performs slightly better than the VIS modality and slightly worse than the real Polar modality. When the VIS modality is replaced by the Polar_G modality, compared with the ResNet and CDCN methods, the ACER value of our method decreases by 0.6% and 1.23%, the accuracy improves by 3.17% and 5.48%, and the F_score improves by 0.39% and 2.5%, respectively. These results demonstrate that the Polar information provided by FC-Net has a significant effect on improving PAD performance, and on the other hand, proves that our FC-Net can generate convincing Polar modalities.

In addition, FAS performance improved after the VIS and Polar_G modalities were fused by PTG-Face, as shown in Table III. PTG-Face achieved the best performance compared to the single-modal methods using either VIS or Polar. In particular, compared to the ResNet (VIS) and CDCN (VIS) methods, PTG-Face shows a 3.20% and 3.77% reduction in ACER, 4.32% and 7.05% improvements in accuracy, and 4.52% and 5.94% improvements in F_score, respectively. These results demonstrate that our approach significantly improves the PAD performance by generating Polar images when only the authentic VIS mode is available as input. An interesting finding is that while the results of the lightweight CDCN network are inferior overall to those of ResNet that have more parameters, PTG-Face with CDCN as the backbone outperforms ResNet, demonstrating that our generated Polar modality provides more discriminative features than the VIS modality. These results highlight the significant improvement

TABLE III
FINE-GRAINED MATERIAL CLASSIFICATION IDENTIFICATION FOR THE CASIA-POLAR DATASET

Method	Metric(%)	Paper	Photo	Screen	3D Mask	Fake Head	Genuine	Overall
ResNet (VIS)	APCER	8.07	11.26	1.58	2.39	12.51	0.00	5.97
	BPCER	0.00	0.00	0.00	0.00	0.00	5.61	0.94
	ACER	4.03	5.63	0.79	1.19	6.26	2.80	3.45
ResNet (Polar)	APCER	11.63	1.94	0.00	1.04	7.37	0.00	3.66
	BPCER	0.00	0.00	0.00	0.00	0.00	0.46	0.08
	ACER	5.82	0.97	**0.00**	0.52	3.69	0.23	1.87
CDCN (VIS)	APCER	18.39	14.01	5.26	4.33	2.67	0.00	7.44
	BPCER	0.00	0.00	0.00	0.00	0.00	3.60	0.60
	ACER	9.19	7.01	2.63	2.16	1.34	1.80	4.02
CDCN (Polar)	APCER	9.74	8.41	0.44	1.20	1.01	0.00	3.47
	BPCER	0.00	0.00	0.00	0.00	0.00	4.29	0.72
	ACER	4.87	4.21	0.22	0.60	0.51	2.14	2.09
Ours (VIS & Polar_G)	APCER	1.23	0.80	0.00	0.77	0.12	0.00	0.48
	BPCER	0.00	0.00	0.00	0.00	0.00	0.11	0.02
	ACER	**0.61**	**0.40**	**0.00**	**0.38**	**0.06**	**0.05**	**0.25**

Fig. 9. The generated polar images in the CASIA-SURF dataset are shown, with the first two rows showing the VIS and generated polar images corresponding to the real face and the last two rows showing the VIS and generated polar images corresponding to the spoofing attack.

in FAS performance via generating Polar images from authentic VIS input only without the use of a polarized camera.

3) FAS Results on Cross-Dataset: The CASIA-SURF public FAS dataset was used to evaluate the generalization of the proposed method to an unknown domain, namely the CASIA-Polar and CASIA-SURF datasets were each used as independent domains in our experiments. We chose CASIA-Polar as the source domain and CASIA-SURF as the unknown domain for testing that was not accessible during the training process.

CASIA-SURF is one of the most widely used datasets in FAS research and consists of three data modalities: VIS, NIR, and depth. CASIA-SURF does not contain Polar data and can demonstrate the superiority of our method in a more objective and significant manner. We first used the FC-Net trained in the CASIA-Polar dataset to generate Polar modes according to the VIS modal data in the CASIA-SURF dataset, and the generated results are shown in Fig. 9.

The performance of the ResNet, CDCN, and PTG-Face methods is then compared by varying the input modalities. By comparing the results of the ResNet and CDCN methods for three modes, namely, VIS, NIR, and Polar_G, that are presented in Table IV, we can see that the Polar_G mode results in better ACER than the other two modes.

TABLE IV
STATISTICS OF THE CASIA-SURF DATASET

Modality	Method	APCER	NPCER	ACER
VIS	ResNet	5.54	16.23	10.88
NIR	ResNet	3.02	2.80	2.91
Polar_G	ResNet	3.82	**1.23**	2.52
VIS	CDCN	4.97	35.29	20.13
NIR	CDCN	6.02	3.76	4.89
Polar_G	CDCN	5.52	3.74	4.63
VIS & NIR	Ours	1.32	1.88	1.60
VIS & Polar_G	**Ours**	**1.33**	1.84	**1.58**

When we replaced the VIS modality with the Polar_G modality, the performance of ACER was reduced by 8.36% and 15.50% in ResNet and CDCN, respectively. This is because the Polar modality contains more discriminative cues that are not apparent in the VIS modality.

In addition, the performance of the dual-stream CDCN was evaluated. With the dual-stream CDCN network, the VIS mode is fused with the NIR and Polar_G modes respectively. Compared to the ResNet and single-mode CDCN methods, the dual-stream CDCN obtained significantly lower ACER results both using VIS & NIR as input and VIS & Polar_G as input, as shown in Table IV. It is noteworthy that the ACER of our method is reduced by 0.02% even when compared to VIS & NIR as input. These findings indicate two important aspects. Firstly, the utilization of dual-stream CDCN for modal fusion can notably enhance the performance of FAS. Secondly, the PTG-Face introduced in this study not only offers PAD features absent in the VIS modality but also effectively utilizes Polar information to augment the learning of the VIS modality.

In summary, the above results show that PTG-Face can provide more recognizable features for FAS tasks. Our algorithm can achieve better FAS performance in FAS systems equipped with only equipped VIS cameras. To further demonstrate the performance of our approach, the results from our tests in real-world scenarios are given in the supplementary material.

4) Comparison With State-of-the-Art Methods: We conducted experiments on the CASIA-Polar dataset to compare our method with state-of-the-art (SOTA) methods. Table V provides comparison results that show that our method significantly outperforms SOTA methods in VIS modality. Compared to the LBP, ResNet, CDCN, and BAS methods, the ACER increased by 6.39%, 3.20%, 4.42%, and 35.05%, respectively, in our method. Additionally, when compared to the SOTA methods in Polar modality, our results are impressive, with the ACER score outperforming the ACER values of the 2D_Polar [15] and PAAS methods by 9.56% and 1.51%, respectively. Moreover, we note that the 2D_Polar method uses statistical methods such as the mean, standard deviation, and kurtosis for FAS; thus, the best results under these three metrics were chosen for comparison. In addition to the comparison with the unimodal method, the performance of our method compared to the conventional multimodal FAS method PSMM-Net [53] when using the VIS mode and generated Polar mode as inputs, the ACER of our method is approximately 1.16% better than that of PSMM-Net. The results in Table V demonstrate that the proposed method learns discriminative features of genuine and fake faces from the VIS and generated Polar images. According to the comparison of the three input modalities, our approach achieved SOTA levels on the ACER, APCER, and BPCER metrics.

TABLE V
COMPARISON OF TEST RESULTS ON THE CASIA-POLAR DATASET

Modality	Method (%)	APCER	BPCER	ACER
VIS	LBP [21]	7.52	5.76	6.64
	ResNet	5.97	0.94	3.45
	CDCN [55]	7.84	1.50	4.67
	BAS [52]	69.52	1.08	35.30
Polar	PAAS [16]	2.79	0.73	1.76
	ResNet	1.95	0.46	1.21
	CDCN [55]	3.41	0.72	2.07
	BAS [52]	31.95	9.98	20.96
	2D_Polar [15]	13.3	6.32	9.81
VIS & Polar_G	PSMM-Net [53]	1.55	1.27	1.41
	Ours	**0.48**	**0.02**	**0.25**

TABLE VI
COMPARISON OF TEST RESULTS FOR THE GENERATION METHODS ON THE CASIA-SURF DATASET

Modality	Method (%)	APCER	BPCER	ACER
VIS & NIR_G	PSMM-Net	2.80	2.10	2.50
	MA-Net [18]	2.40	**1.70**	2.00
VIS & Polar_G	PSMM-Net	2.01	1.94	1.97
	Ours	**1.33**	1.84	**1.58**

In Table VI, we show the comparison between the results of our method and the results of the method described in [18] under the CASIA-SURF dataset. The ACER result for PSMM-Net is 0.53% lower in the Polar_G modality than in NIR_G. The ACER result for PTG-Face is 0.42% lower than those for MA-Net.

To better analyze our proposed method, we compared it with the SOTA NIR generation method [18] in CASIA-SURF, where all experiments were performed on the setup in [18]. Compared to [18], lower ACER results were achieved using our Polar_G modal in PSMM-Net than the NIR modal generated in [18], with an ACER reduction of 0.53%. Meanwhile, the ACER result for PTG-Face is 0.42% lower than that of MA-Net. Thus, our method and the Polar_G modes outperform the SOTA methods.

V. Conclusion

In this paper, we revisit the application of polar patterns in FAS missions. We propose PTG-Face, a new FAS method that uses physical cues from VIS and Polar images in the FAS task without the need for additional polarized imaging equipment. We design an FC-Net and propose a novel frequency domain consistency loss that translates VIS images to Polar images based on the collected VIS images to obtain discriminable Polar genuine and false face images. Then, we use a dual-stream CDCN model to learn and extract features from the real VIS face images and the generated Polar face images for genuine-fake face classification via feature fusion. Extensive experimental results show that our approach can not only generate realistic Polar face images, but also tap the intrinsic features of genuine and fake faces, and achieve excellent results in FAS classification performance. Our planned future work includes 1) optimizing the Polar generation network by in-depth analysis of Polar features, and 2) establishing a more suitable Polar-based FAS benchmark dataset.

References

[1] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and Fisher vector encoding," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 141–145, Feb. 2017.

[2] L.-B. Zhang, F. Peng, L. Qin, and M. Long, "Face spoofing detection based on color texture Markov feature and support vector machine recursive feature elimination," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 56–69, Feb. 2018.

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

[4] X. Tu, H. Zhang, M. Xie, Y. Luo, Y. Zhang, and Z. Ma, "Enhance the motion cues for face anti-spoofing using CNN-LSTM architecture," 2019, *arXiv:1901.05635*.

[5] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, "Detection of face spoofing using visual dynamics," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 762–777, Apr. 2015.

[6] M. Asim, Z. Ming, and M. Y. Javed, "CNN based spatio-temporal feature extraction for face anti-spoofing," in *Proc. 2nd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2017, pp. 234–238.

[7] M. Sajjad et al., "CNN-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognit. Lett.*, vol. 126, pp. 123–131, Sep. 2019.

[8] A. Liu et al., "Cross-ethnicity face anti-spoofing recognition challenge: A review," *IET Biometrics*, vol. 10, no. 1, pp. 24–43, Jan. 2021.

[9] K. Kotwal et al., "Domain-specific adaptation of CNN for detecting face presentation attacks in NIR," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 1, pp. 135–147, Jan. 2022.

[10] J. Guo, X. Zhu, J. Xiao, Z. Lei, G. Wan, and S. Z. Li, "Improving face anti-spoofing by 3D virtual synthesis," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–8.

[11] A. Liu and Y. Liang, "MA-ViT: Modality-agnostic vision transformers for face anti-spoofing," 2023, *arXiv:2304.07549*.

[12] Z. Yu, A. Liu, C. Zhao, K. H. M. Cheng, X. Cheng, and G. Zhao, "Flexible-modal face anti-spoofing: A benchmark," 2022, *arXiv:2202.08192*.

[13] A. Kimachi, "Polarization imaging for material classification," *Opt. Eng.*, vol. 47, no. 12, Dec. 2008, Art. no. 123201.

[14] E. M. Rudd, M. Günther, and T. E. Boult, "PARAPH: Presentation attack rejection by analyzing polarization hypotheses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 171–178.

[15] A. Z. A. Aziz, H. Wei, and J. Ferryman, "Face anti-spoofing countermeasure: Efficient 2D materials classification using polarization imaging," in *Proc. 5th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2017, pp. 1–6.

[16] Y. Tian, K. Zhang, L. Wang, and Z. Sun, "Face anti-spoofing by learning polarization cues in a real-world scenario," in *Proc. 4th Int. Conf. Adv. Image Process.*, Nov. 2020, pp. 129–137.

[17] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 319–328.

[18] A. Liu et al., "Face anti-spoofing via adversarial cross-modality translation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2759–2772, 2021.

[19] F. Jiang, P. Liu, X. Shao, and X. Zhou, "Face anti-spoofing with generated near-infrared images," *Multimedia Tools Appl.*, vol. 79, nos. 29–30, pp. 21299–21323, Aug. 2020.

[20] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–8.

[21] T. D. F. Pereira, A. Anjos, J. M. D. Martino, and S. Marcel, "LBP—TOP based countermeasure against face spoofing attacks," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2012, pp. 121–132.

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

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

[24] Z. Xu, S. Li, and W. Deng, "Learning temporal features using LSTM-CNN architecture for face anti-spoofing," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 141–145.

[25] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Dec. 2016, pp. 1–6.

[26] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral SWIR imaging," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[27] A. Gumaei, R. Sammouda, A. M. S. Al-Salman, and A. Alsanad, "Anti-spoofing cloud-based multi-spectral biometric identification system for enterprise security and privacy-preservation," *J. Parallel Distrib. Comput.*, vol. 124, pp. 27–40, Feb. 2019.

[28] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 42–55, 2020.

[29] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, and S. Marcel, "Deep models and shortwave infrared information to detect face presentation attacks," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 4, pp. 399–409, Oct. 2020.

[30] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Face liveness detection by learning multispectral reflectance distributions," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 436–441.

[31] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3D mask face anti-spoofing with remote photoplethysmography," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 85–100.

[32] C. Yao et al., "rPPG-based spoofing detection for face mask attack using efficientnet on weighted spatial–temporal representation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3872–3876.

[33] J. Hernandez-Ortega, J. Fierrez, A. Morales, and P. Tome, "Time analysis of pulse-based face anti-spoofing in visible and NIR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 544–552.

[34] Z. Wang et al., "Exploiting temporal and depth information for multi-frame face anti-spoofing," 2018, *arXiv:1811.05118*.

[35] S. Kim, Y. Ban, and S. Lee, "Face liveness detection using a light field camera," *Sensors*, vol. 14, no. 12, pp. 22471–22499, Nov. 2014.

[36] M. Liu et al., "Light field-based face liveness detection with convolutional neural networks," *J. Electron. Imag.*, vol. 28, no. 1, 2019, Art. no. 013003.

[37] S. Sun, Y. Tian, Y. Tang, and B. Wu, "Anti-spoofing face recognition using infrared structure light," in *Frontiers in Optics*. Washington, DC, USA: Optica Publishing Group, 2020, pp. 1–3.

[38] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves, "Conditional image generation with PixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–12.

[39] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[42] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.
[43] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.
[44] K. Panetta et al., "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.
[45] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1522–1535, Oct. 2012.
[46] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
[47] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13899–13909.
[48] L. B. Wolff and T. E. Boult, "Constraining object features using a polarization reflectance model," *Phys.-Based Vis., Princ. Pract., Radiometry*, vol. 1, p. 167, Jan. 1993.
[49] J. F. de Boer and T. E. Milner, "Review of polarization sensitive optical coherence tomography and Stokes vector determination," *J. Biomed. Opt.*, vol. 7, no. 3, pp. 359–371, 2002.
[50] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[51] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 612–618.
[52] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 389–398.
[53] S. Zhang et al., "CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 2, pp. 182–193, Apr. 2020.
[54] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1178–1186.
[55] Z. Yu et al., "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5294–5304.
[56] Z. Yu et al., "Multi-modal face anti-spoofing based on central difference networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2766–2774.
[57] A. Liu et al., "Contrastive context-aware learning for 3D high-fidelity mask face presentation attack detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2497–2507, 2022.
[58] S. Elliott, *Biometrics International Standards*, Standard JTC 1 SC 37, Biometrics Standards, Perform., Assurance Lab., Purdue Univ., West Lafayette, IN, USA, 2002.
[59] Z.-Q. John Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

Yu Tian received the B.E. and M.S. degrees from the North University of China, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the School of Optics and Photonics, Beijing Institute of Technology. He is also an Engineer with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, computational photography, and biometric recognition.

Yalin Huang received the B.E. degree from the College of Information Science and Engineering, Ocean University of China, in 2017. He is currently pursuing the master's degree with the Institute of Automation, Chinese Academy of Sciences. His research interests include machine vision, deep learning, and face anti-spoofing.

Kunbo Zhang (Member, IEEE) received the B.E. degree in automation from the Beijing Institute of Technology in 2006 and the M.Sc. and Ph.D. degrees in mechanical engineering from the State University of New York at Stony Brook, Stony Brook, NY, USA, in 2008 and 2011, respectively. From 2011 to 2016, he was with the Advanced Manufacturing Engineering Group, Nexteer Automotive, Saginaw, MI, USA. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China. His research interests include computational photography, biometric imaging, machine vision, and intelligent systems.

Yue Liu (Member, IEEE) received the M.Sc. and Ph.D. degrees from Jilin University, China, in 1996 and 2000, respectively. He was a Visiting Research at relevant laboratories, such as the Harvard Medical School; the University of California, Berkeley; the Georgia Institute of Technology; Temple University; and Australian National University, USA. He is currently a Professor with the Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, China. He has published more than 100 papers in international journals and conferences. His research interests include virtual and augmented reality, human–computer interaction, and computer vision.

Zhenan Sun (Senior Member, IEEE) received the B.E. degree in industrial automation from the Dalian University of Technology, China, in 1999, the M.S. degree in system engineering from the Huazhong University of Science and Technology, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, China, in 2006. He is currently a Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. He is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, China. He has authored/coauthored more than 200 technical articles. His research interests include biometrics, pattern recognition, and computer vision. He is the Vice President of the Technical Committee on Biometrics and the International Association for Pattern Recognition (IAPR), and an IAPR Fellow. He serves as an Associate Editor for the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE (T-BIOM) and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS).