Data-Fusion-Based Two-Stage Cascade Framework for Multimodality Face Anti-Spoofing

Weihua Liu, Xiaokang Wei, Tao Lei, *Senior Member, IEEE*, Xingwu Wang, Hongying Meng, *Senior Member, IEEE*, and Asoke K. Nandi, *Fellow, IEEE*

***Abstract*—Existing face anti-spoofing models using deep learning for multimodality data suffer from low generalization in the case of using variety of presentation attacks, such as 2-D printing and high-precision 3-D face masks. One of the main reasons is that the nonlinearity of multispectral information used to preserve the intrinsic attributes between a real and a fake face is not well extracted. To address this issue, we propose a multimodility data-based two-stage cascade framework for face anti-spoofing. The proposed framework has two advantages. First, we design a two-stage cascade architecture that can selectively fuse low-level and high-level features from different modalities to improve feature representation. Second, we use multimodality data to construct a distance-free spectral on RGB and infrared to augment the nonlinearity of data. The presented data fusion strategy is different from popular fusion approaches, since it can strengthen discrimination ability of network models on physical attribute features than identity structure features under certain constraints. In addition, a multiscale patch-based weighted fine-tuning strategy is designed to learn each specific local face region. The experimental results show that the proposed framework achieves better performance than other state-of-the-art methods on both benchmark data sets and self-established data sets, especially on multimaterial masks spoofing.**

***Index Terms*—Convolutional neural network (CNN), deep learning, face anti-spoofing, multimodality fusion.**

Manuscript received May 30, 2020; revised October 26, 2020 and January 5, 2021; accepted February 27, 2021. Date of publication March 8, 2021; date of current version June 10, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61871259, Grant 61811530325 (IEC\NSFC\170396, Royal Society, U.K.), Grant 61871260, Grant 61672333, and Grant 61873155; in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JM6065; and in part by the Science and Technology Program of Shaanxi Province under Grant 2020NY-172. *(Corresponding author: Tao Lei.)*

Weihua Liu is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China, and also with the Laboratory of Intelligent Image Processing, Orbbec Company, Shenzhen 518058, China (e-mail: liuweihua@orbbec.com).

Xiaokang Wei is with the Laboratory of Intelligent Image Processing, Orbbec Company, Shenzhen 518058, China. (e-mail: weixiaokangxj@163.com).

Tao Lei and Xingwu Wang are with the Shaanxi Joint Laboratory of Artificial Intelligence and the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: leitao@sust.edu.cn; wangxwu1949@163.com).

Hongying Meng is with the Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, U.K. (e-mail: hongying.meng@brunel.ac.uk).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, U.K., and also with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: asoke.nandi@brunel.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2021.3064679.

Digital Object Identifier 10.1109/TCDS.2021.3064679

I. Introduction

HUMAN face is one of the most salient and stable biometrics, it often relies on various kinds of interactive AI systems, and has been widely used in many crowd gathering and sensitive areas [1]–[3], such as attendance registration, security surveillance, etc. In spite of successful applications in many types of face authentication scenarios, most of existing face recognition systems are easily spoofed by presentation attacks (PAs) ranging from a 2-D printing attack or a vivid 3-D-mask attack [4], [5]. For example, with the help of silicone or latex masks, users easily portray another identity or obfuscate their identity for entertainment purposes. However, such masks have been treated as criminal tools to deceive automatic face recognition systems. Therefore, it is important to distinguish a real face and a fake face for face recognition and authentications systems. In general, a robust face recognition system can cope with variants of face states, such as face partial occlusion, the change of face expression, etc. On the contrary, variant face presentations should be strictly restricted on face anti-spoofing tasks, and an entire frontal face presentation is required. More importantly, an advanced face anti-spoofing model needs to show strong discriminability on intradata set with prior defined face knowledge, and performs well on interdata set with unknown faces.

Most popular face anti-spoofing methods extract generalized likeness feature under binary supervision. In practice, a well discriminative feature map is composed of structure clues, texture clues, depth clues, material clues, etc. There are many structure and texture distinctions between original and recaptured images. Here, we present two obvious distinctions. The first is light reflection. Typically, fake face materials are much flat and smooth than real faces, easily cause specular reflection, especially under active infrared (IR) light spectral. Second, Moiré Pattern-based image is formed due to superimposing of the gratings and can be extracted by traditional feature descriptors, such as local binary pattern (LBP), histograms of oriented gradients (HOG), difference of Gaussian (DOG), speeded up robust features (SURF), etc. In addition, face anti-spoofing models are also improved by considering the influence from image blur, distortion and noise, etc.

Under the help of depth sensors, such as time-of-flight, structure light, stereo cameras, etc., point clouds of objects can be directly constructed to prevent 2-D-based fake face attacks, such as flatten screens, papers, etc. Most existing studies resort to a face detector to obtain face landmarks, and

then reconstruct 3-D face models based on the stereo vision. Motivated by these ideas, many auxiliary depth supervised face anti-spoofing models have been developed. Intuitively, the face-like depth can be regressed from images of real faces, whereas the none face depth is regressed from images of fake faces, such as printing style, replaying style, etc.

By providing richer spectral information, spectral imaging is far beyond human visual perception ability in the field of object detection and recognition. The researchers manifest that compared to RGB or monochrome cameras, the utilization of multispectral imaging can enhance spatial heterogeneity that is not easily captured by the human visual system, and thus leads to better face detection and recognition. Clearly, the spectral signature between real and fake faces provides additional spectral–spatial information that is helpful for improving face anti-spoofing.

Although some significant models are proposed and used for face anti-spoofing as aforementioned, most of them are trained on 2-D data sets. Therefore, these models are still vulnerable in 3-D reality environments. To address this issue, we propose a robust face anti-spoofing model that focuses on defending against both 2-D and 3-D face PAs. The contributions of our work are given as follows.

1) We design a two-stage cascade framework that takes advantage of depth, color, and IR data streams to hierarchically capture discriminative details on 3-D structure (i.e., depth) clues and intrinsic face (i.e., RGB and IR) properties.
2) We construct a distance-free image from multimodality data and develop an effective data fusion strategy by mixing the distance-free image, RGB, and IR, to enlarge the feature difference between real and fake faces.
3) We establish a multiimage patterns data set (MIPD) containing variant 2-D and 3-D PAs; hence can be used as a benchmark data set for extensively verifying the generalization ability of other face anti-spoofing models.

II. Related Work

Popular face anti-spoofing models are designed in accordance with several face spoofing levels, which can be mainly categorized into four groups: 1) monocular color images; 2) depth images; 3) multispectral images; and 4) multimodality images.

A. Face Anti-Spoofing

Monocular Color Images: As the most face spoofing PAs depend on image recapturing or portable electronic devices presentation [6], many face anti-spoofing algorithms pay much attention on image quality, image texture, etc., [7]–[9]. Yang *et al.* [10] introduced a component dependent descriptor by partitioning a face to six parts and each part was represented by using the bag of words model. However, the framework requires a strong face detection module to sustain its performance. Actually, it is a challenging task to guide against spoofs by only considering static images in 2-D space. Therefore, video-based face anti-spoofing models are gradually presented by taking the advantage of temporal information [11], [12]. Early studies allow the user to interact in front of devices, such as blinking eyes, nodding head, etc., but the interactions reduce the quality of experiences. Siddiqui *et al.* [13] used temporal evidence aggregation over face region and scene of video images, which performs well on 2-D face attack data sets due to the synthesized multifeatures, but fails to judge the 3-D realistic masks. In addition, long short-term memory (LSTM) [14] can recurrently learn features to obtain context information, but it suffers from the heavy computational burden. To obtain intrinsic liveness cues, researchers propose a remote photo plethysmography (rPPG) technique [15] to detect the heartbeat signal from face appearance. This technique can detect the blood flow based on the absorption and reflection of light passing through human skin. Hence, fake faces can be recognized by identifying such subtle blood color variations. However, rPPG shows weak robustness in real scenarios because of environmental noise, such as camera motion, dim light or image low resolution, etc.

Depth Images: The above algorithms, whether based on static images or clips of videos, are vulnerable to 3-D fake attacks [4]. For this problem, a 3-D face reconstruction is useful by dramatically rejecting a bunch of fake faces from electronic devices or flatten papers. However, the algorithms of 3-D face reconstruction are seldom studied since that both the intrinsic and extrinsic parameters of visible light cameras are hard to be acquired precisely. In view of this, Wang *et al.* [16] proposed a sparse 3-D structure recovered method by capturing at least two face images from different viewpoints. This method outperforms other texture-based methods, especially on fake faces from printed photos. Unfortunately, it is not very convincing that their fake samples are all warped papers, rather than 3-D face masks. Recently, researchers find that the depth estimation technology is beneficial for modeling face anti-spoofing by utilizing face depth map as a supervised signal. Compared to the binary cross-entropy loss that easily learns the arbitrary patterns, depth supervision can learn better spoofing patterns. For example, Yu *et al.* [17] considered pseudoface depth maps as the auxiliary supervised signal, and proposed a novel operation called central difference convolution (CDC) for face anti-spoofing, which is able to capture intrinsic detailed patterns via aggregating both intensity and gradient information.

Multispectral Images: With the demands of upgrading the security level in practical applications, the multispectral technique is used to evaluate the essential attribute of fake and real faces. Compared to general RGB cameras that only provide three channels from visible wavelength bands, multispectral imagers provide more broadly wavelength bands, which can greatly resist on the interference of illumination and reflect more intrinsic object attributes [18]–[19]. Typically, reflectance images of real faces show low intensity and uniform distribution while reflectance images of 2-D/3-D spoofing faces show high intensity and nonuniform distribution [20], [21]. In the past few years, few studies have been devoted to multispectral-based face anti-spoofing due to the lack of portable devices. In one of most representative studies, Zhang *et al.* [22] conducted a complete experiment with wavelengths span from visible to near IR on face anti-spoofing. In their experiments, the albedo curve of different materials and skin colors is presented and some clear comparisons are

shown to verify the effectiveness of the multispectral technique for face anti-spoofing.

Multimodality Images: Multimodality methods provide more robust performance than the above single modality techniques since they utilize complementary information among different modality data [23]. Based on the assumption, Zhang *et al.* [25] released a large scale of multimodality data set (CASIA-SURF) for face anti-spoofing, which consists of 1000 subjects with 21 000 videos, and each sample contains three modality data (i.e., RGB, depth, and IR). With CASIA-SURF data set, Parkin and Grinchuk [26] proposed a multilevel feature aggregation network that achieves feature fusion of different modality data at both coarse and fine levels. Also, Shen *et al.* [27] proposed a patch-based multistream fusion convolutional neural network (CNN) architecture to extract local-spoof discriminative information. More recently, Liu *et al.* [24] released the CASIA-SURF cross-ethnicity face anti-spoofing (CeFA) data set, covering 3 ethnicities, 3 modalities, 1607 subjects, and 2-D plus 3-D attacks. They also proposed a novel static-dynamic fusion mechanism as a strong baseline to learn complementary information from multiple modalities. To further differentiate feature extraction among different modality data, Yang *et al.* [28] designed a selective modal pipeline of the fusion network for multimodal face anti-spoofing. However, the aforementioned works are evaluated on a small face anti-spoofing data set with limited subjects and samples, which easily leads to the problem of over-fitting.

B. Deep Learning for Face Anti-Spoofing

In Section II-A, monocular color-based methods are overemphasizing on optical imaging difference between real and fake faces. However, with the rapid development of printing techniques, false positive detections of such methods are gradually increasing and make them much more unreliabe.

For deep learning, a large number of CNNs achieve remarkable success in various tasks of object detection and recognition, [17], [29]–[32]. As for face anti-spoofing, it can be treated as a special case of image classification tasks. It aims to distinguish between real and fake faces by outputting a binary prediction [33], [34]. Rehman *et al.* [35] introduced a continuous data-randomization mechanism in the form of small minibatches to train several benchmark CNNs, which avoid overfitting on data sets with limited fake samples. However, it is insufficient to only take feature maps from the last layer of benchmark CNNs (VGG, ResNet, etc.) as the output on face anti-spoofing tasks, multilevel feature maps containing both texture and reflection clues from hierarchical layers are expected. Alotaibi and Mahmood [36] analyzed 2-D images using nonlinear diffusion that is helpful for distinguishing boundaries between real and fake faces. After that, these diffused images are taken as the input of a deep CNN to extract useful high-level features. However, diffused features are greatly influenced by original image quality, which reduces the final face anti-spoofing performance. Feng *et al.* [37] proposed a hierarchical neural network for face anti-spoofing by combining image quality cues and motion cues. Although the network achieves better performance than any of static image cues, the image sequence is required and the computational cost is expensive. Following the similar idea, a method using the combination of two-stream CNN is proposed to extract local features and holistic depth maps from face images [38]. These methods are clearly superior to conventional feature descriptors.

III. Methodology

A. System Architecture

In this section, we propose an end-to-end strategy to extract the fusion information from aforementioned data sources by designing a two-stage cascade network. Specifically, two convolutional networks with similar architecture are designed and a cascade framework for face anti-spoofing is presented. For the first stage of the framework, the multipreprocessed depth faces as the input of D-Net are employed to discriminate significant spoofing attacks, such as printed photos, replayed videos, etc. Before the training, depth faces are preprocessed by three ways, including depth normalization, depth face scale embedding, and normal orientation embedding. For our design, if the predication score from this model is greater than 0.5, the model output is "fake face." Otherwise, the second stage, namely, M-Net is implemented. In the M-Net stage, the fusion multimodality formation of RGB and IR, i.e., stack, summation, and difference, is fed into M-Net for further resisting more 3-D mask attacks. The pipeline of our proposed framework is shown in Fig. 1.

B. Depth-Image Preprocessing and Multimodality Fusion

Before implementing the face anti-spoofing model, the compact face area of IR, depth, and RGB should be synchronously extracted by utilizing the face-detection module [39]. For popular face anti-spoofing systems, on the one hand, frontal face images (within 206° of pitch, yaw, and roll) are required because they can provide richer depth information than side faces. On the other hand, the subsequent face recognition module also needs frontal face images for feature representation. Based on this consideration, here we remove exaggerated side faces by rejecting the large angle between face norm and rolling axis of sensors.

1) Depth Image Processing (Depth Face Normalization): Due to the illumination reflection and self-occlusion, a raw depth face often includes some noisy holes leading to the miss of part depth information. To address this issue, the depth completion is first used to recover an entire face depth information to achieve a smooth and continuous face depth surface; second, the depth connected component [40] is further used to suppress noise from background. Specifically, the largest depth connection area is extracted as the face profile region and the value of background pixels is set to a constant 255. After that, the current preprocessed raw depth face, denoted as $\mathbf{I}^o$, is finely normalized by globally computing the maximum and minimum depth values within face region, denoted by $d_o^{\max}$ and $d_o^{\min}$, respectively. The normalized depth face $\mathbf{D}^n$ is defined as

$$\mathbf{D}^n = \text{floor}\left(\frac{\mathbf{I}^o - \mathbf{K}^{m\times n} \times d_o^{\min}}{d_o^{\max} - d_o^{\min}} \times 255\right) \tag{1}$$

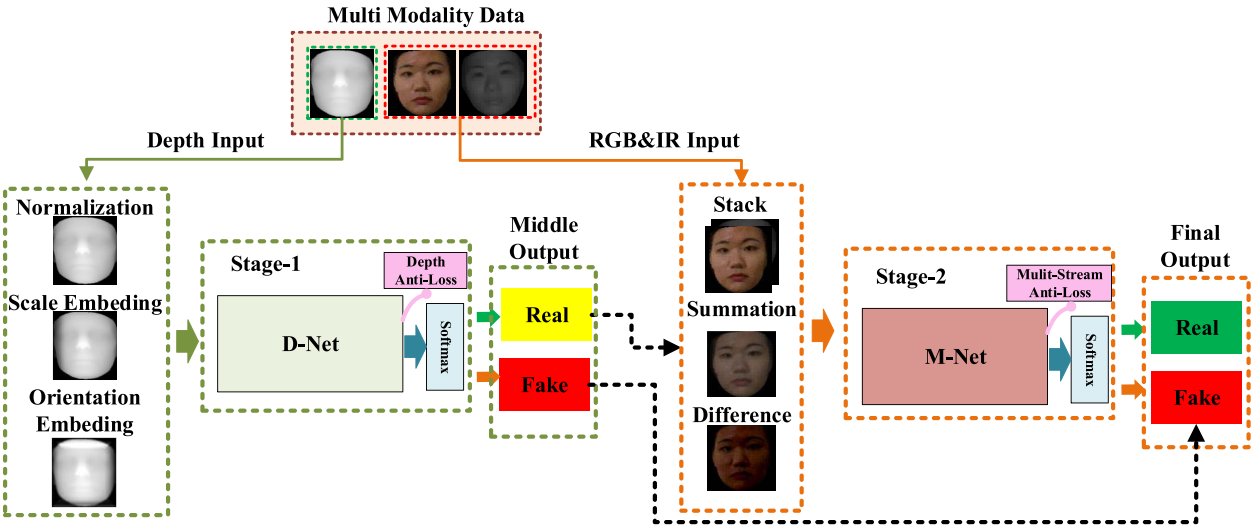


Fig. 1. Overall cascade framework for face anti-spoofing, where the D-Net takes depth images as the network input and the M-Net takes fused images as the network input.

where $\mathbf{K}^{m\times n}$ is a matrix and the values of all elements in $\mathbf{K}^{m\times n}$ are 1. Floor is the function that takes as input a real number and gives as output the greatest integer less than or equal to that real number.

Depth Face Scale Embedding: Generally, human face sizes are maintained at a certain level and are proportional to camera distances, which can roughly reject abnormal faces at oversized scales. Instead of setting a proper threshold to exclude faces with abnormal scales, the face size information is embedded pixelwise into normalized face $\mathbf{D}^s$

$$\mathbf{D}^s = \text{floor}\left(\frac{\mathbf{I}^s - \mathbf{K}^{m\times n} \times d_s^{\min}}{d_s^{\max} - d_s^{\min}} \times 255\right) \tag{2}$$

where $\mathbf{I}^s$ represents the scale matrix of a depth face and $\mathbf{I}^s = \mathbf{I}^o/s$. The face area size $s = w \times h$ can be obtained in terms of face bounding box (p_x, p_y, w, h). Similarly, $d_s^{\min}$ and $d_s^{\max}$ represent the minimum and maximum value of $\mathbf{I}^s$, respectively.

Normal Orientation Embedding: Here, we calculate the normal vector for each pixel in the raw depth face $\mathbf{I}^o$, and it is denoted by $\vec{v}^{\text{norm}}$. As aforementioned, five key face landmarks provided by the face detector, are taken as initial points to fit a face plane τ. Hence, the face planner normal vector $\vec{v}_\tau^{\text{norm}}$ is orthogonal to the plane τ. After that, a matrix $\mathbf{I}^r$ can be formed by computing arc-cosine between $\vec{v}_\tau^{\text{norm}}$ and each $\vec{v}^{\text{norm}}$. Consequently, the raw depth face of $\mathbf{I}^o$ is replaced by the normalized face $\mathbf{D}^r$

$$\mathbf{D}^r = \text{floor}\left(\frac{\mathbf{I}^r - \mathbf{K}^{m\times n} \times d_r^{\min}}{d_r^{\max} - d_r^{\min}} \times 255\right) \tag{3}$$

where $d_r^{\min}$ and $d_r^{\max}$ represent the minimum and maximum values of $\mathbf{I}^r$, respectively. According to (3), the orientation information on each face is preserved and can be further extracted in the subsequent CNN modules.

2) Multispectral Image Processing (Reflectance Analysis): From the perspective of spectroscopy, a real face and a 3-D face mask can be distinguished by computing their spectral reflectance due to different materials. According to the Lambertian reflectance model [41], the reflectance light intensity $I(p)$ of pixel p can be written as

$$I(p) = s \times \rho \times \cos(\theta) \tag{4}$$

where s is the external light intensity, ρ is the object albedo at pixel p related to physical properties of the object material at each wavelength ω, and θ is the angle between pixel's normal and receiver's viewpoint.

According to the Beer–Lambert law, the attenuation of light through the air is $s = s_0 e^{-cd}$, where s_0 is the light source intensity, c is the attenuation coefficient in the air, and d is the distance traveled. For simplicity, we denote $s = s_0 e^{-cd}$ as a function $D(d)$, which is a monotone decreasing function of distance d, and by jointing (4), we have

$$I(p) = D(d) \times \rho \times \cos(\theta). \tag{5}$$

From (5), it can be found that $I(p)$ is linear with ρ under the same $\cos(\theta)$ and external light intensity. This makes it possible to tackle the face anti-spoofing problem because the indistinguishable fake faces exhibit quite different albedo properties under the condition of multispectrum. If we define a real face f_1 and a fake face f_2, respectively, $\rho_{f_1} \neq \rho_{f_2}$ under most of the wavelengths due to different materials. Specifically, giving a wavelength ω_1, from (5), we have

$$I^{\text{face},\omega_1} = D(d_1)\rho^{\text{face},\omega_1}\cos(\theta) \tag{6}$$

$$I^{\text{spoof},\omega_1} = D(d_2)\rho^{\text{spoof},\omega_1}\cos(\theta). \tag{7}$$

According to (6) and (7), f_1 and f_2 are undistinguishable since there is always a proper distance pair d_1 and d_2 (d_1 is unnecessarily equivalent to d_2) such that

$$I^{\text{face},\omega_1} = I^{\text{spoof},\omega_1} \rightarrow \frac{\rho^{\text{face},\omega_1}}{\rho^{\text{spoof},\omega_1}} = \frac{D(d_2)}{D(d_1)}. \tag{8}$$

To solve the problem imposed by distance, a multispectral solution is considered. If another proper wavelength ω_2 is selected, then we have

$$I^{\text{face},\omega_2} \neq I^{\text{spoof},\omega_2} \rightarrow \frac{\rho^{\text{face},\omega_2}}{\rho^{\text{spoof},\omega_2}} \neq \frac{D(d_2)}{D(d_1)}. \tag{9}$$

From above equations, we clearly find that if both ω_1 and ω_2 are well selected, and

$$\frac{\rho^{\text{spoof},\omega_2}}{\rho^{\text{face},\omega_2}} \neq \frac{\rho^{\text{spoof},\omega_1}}{\rho^{\text{face},\omega_1}} \tag{10}$$

then real and fake faces can be easily distinguished without any distance requirement.

Nonlinear Data Fusion: From the above section, the selection of proper wavelengths is crucial for improving the classification of real-fake faces when there is no limitation on distance between faces and sensors. However, it is difficult to select suitable values of ω_1 and ω_2 for different tasks. Thanks to the registered image between RGB and IR, we consider to establish a distance free image

$$I^{\omega}_{abd} = \frac{I^{\omega}(p)}{D(p)} \tag{11}$$

where $I^{\omega}(p)$ is the reflectance light intensity at pixel p in (5). As mentioned above, $D(p)$ is a monotone decreasing function of distance d. Furthermore, by combining $I^{\text{face},\omega_2}_{abd}$ and $I^{\text{spoof},\omega_2}_{abd}$ with I^{face,ω_1} and $I^{\text{spoof},\omega_1}$, respectively, the nonlinearity between two spectral data is enhanced as follows:

$$\begin{aligned}\frac{I^{\text{face},(\omega_1,\omega_2)}_{\text{sum}}}{I^{\text{spoof},(\omega_1,\omega_2)}_{\text{sum}}} &= \frac{I^{\text{face},\omega_1} + I^{\text{face},\omega_2}_{abd}}{I^{\text{spoof},\omega_1} + I^{\text{spoof},\omega_2}_{abd}} \\ &\to \frac{\rho^{\text{face},\omega_1} D(d_1) + \rho^{\text{face},\omega_2}}{\rho^{\text{spoof},\omega_1} D(d_2) + \rho^{\text{spoof},\omega_2}} \end{aligned} \tag{12}$$

$$\begin{aligned}\frac{I^{\text{face},(\omega_1,\omega_2)}_{\text{diff}}}{I^{\text{spoof},(\omega_1,\omega_2)}_{\text{diff}}} &= \frac{I^{\text{face},\omega_1} - I^{\text{face},\omega_2}_{abd}}{I^{\text{spoof},\omega_1} - I^{\text{spoof},\omega_2}_{abd}} \\ &\to \frac{\rho^{\text{face},\omega_1} D(d_1) - \rho^{\text{face},\omega_2}}{\rho^{\text{spoof},\omega_1} D(d_2) - \rho^{\text{spoof},\omega_2}} \end{aligned} \tag{13}$$

where the factors of $\cos(\theta)$ and s_0 are removed as mentioned in (8). According to (12) and (13) and the fundamental assumption $\rho^{\text{face},\omega_1} \neq \rho^{\text{spoof},\omega_1}$ and $\rho^{\text{face},\omega_2} \neq \rho^{\text{spoof},\omega_2}$, we get the following inequality without distance constraint:

$$\frac{\rho^{\text{face},\omega_1} D(d_1) + \rho^{\text{face},\omega_2}}{\rho^{\text{spoof},\omega_1} D(d_2) + \rho^{\text{spoof},\omega_2}} \neq \frac{\rho^{\text{face},\omega_1} D(d_1) - \rho^{\text{face},\omega_2}}{\rho^{\text{spoof},\omega_1} D(d_2) - \rho^{\text{spoof},\omega_2}}. \tag{14}$$

In general, we designed three image fusion ways to strength the albedo nonlinearity between real and fake faces.

Data Stack: A four channel multispectral data structure is established by parallelly stacking the albedo RGB image I^{RGB}_{abd} and albedo IR image $\mathbf{I}^{ir}_{abd}$

$$\mathbf{I}^{\text{stack}} = \left[\mathbf{I}^{r}_{abd}, \mathbf{I}^{g}_{abd}, \mathbf{I}^{b}_{abd}, \mathbf{I}^{ir}_{abd}\right] \tag{15}$$

where $\mathbf{I}^{\text{RGB}}_{abd} = [\mathbf{I}^{r}_{abd}, \mathbf{I}^{g}_{abd}, \mathbf{I}^{b}_{abd}]$.

Data Summation: The aligned images of RGB and IR are summed together, i.e.,

$$\mathbf{I}^{\text{sum}} = \left[\mathbf{I}^{r} + \mathbf{I}^{ir}_{abd}, \mathbf{I}^{g} + \mathbf{I}^{ir}_{abd}, \mathbf{I}^{b} + \mathbf{I}^{ir}_{abd}\right] \tag{16}$$

then $\mathbf{I}^{\text{sum}}$ is normalized and the value of each pixel ranges from 0 to 255.

Data Difference: Aligned images of RGB and IR are subtracted

$$\mathbf{I}^{\text{diff}} = \left[\left\|\mathbf{I}^{r} - \mathbf{I}^{ir}_{abd}\right\|, \left\|\mathbf{I}^{g} - \mathbf{I}^{ir}_{abd}\right\|, \left\|\mathbf{I}^{b} - \mathbf{I}^{ir}_{abd}\right\|\right]. \tag{17}$$

However, since the subtraction easily leads to the reduction of image intensity contrast, we use gamma correction to enhance the reduction

$$\mathbf{I}^{\text{diff}}_{\gamma} = f\left(\mathbf{I}^{\text{diff}}\right) \tag{18}$$

where the parameter γ is set to $1/2.2$ empirically.

3) Data Augmentation: After observing the difference between real and fake faces, we can intuitively conclude that details of 3-D fake faces, such as the jumping edge and the texture structure, are finer and smoother than real faces. To effectively preserve such local consistency and enhance M-Net perception ability on saliency spoofing features, we use the Fmix method [42] to achieve data augmentation. Specifically, we implement inverse Discrete Fourier transforming (IDFT) on a low pass filter $\mathbf{f}[\mathbf{Z}]$ to obtain a grayscale image $\mathbf{Z}^G = \Re(F^{-1}(\mathbf{f}[\mathbf{Z}]))$, where $\mathbf{Z}$ denotes a complex random variable with density $\mathbb{N}(\mathbf{0}, \mathbf{I}_{w\times h})$, $\mathbf{Z} \in \mathbb{C}^{w\times h}$. Afterward, we can get the mask image $\mathbf{m}^{\lambda}$ by setting the top $T_s = \lambda \times w \times h$ elements of $\mathbf{Z}^G$

$$\mathbf{m}^{\lambda}\left(\mathbf{Z}^G_{i,j}\right) = \begin{cases} 1, & \text{if } \mathbf{Z}^G_{i,j} \in \text{top}\left(T_s, \mathbf{Z}^G\right) \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

where the parameter λ is set to 0.8, empirically. According to the above mask image $\mathbf{m}^{\lambda}$, e.g., the fusion of $\mathbf{I}^{\text{sum}}$ and $\mathbf{I}^{\text{diff}}_{\gamma}$ is defined as

$$\mathbf{I}^{\text{fuse}}_{\text{aug}} = \mathbf{m}^{\lambda} \odot \mathbf{I}^{\text{sum}} + \left(1 - \mathbf{m}^{\lambda}\right) \odot \mathbf{I}^{\text{diff}}_{\gamma}. \tag{20}$$

According to (20), the data nonlinearity is further enhanced by fusing diverse image regions between $\mathbf{I}^{\text{diff}}$ and $\mathbf{I}^{\text{sum}}$. Therefore, the proposed data augmentation strategy is useful for improving feature representation of our model.

C. Design of Network

To obtain better feature representation, we design a new network architecture that consists of three branches, where ResNeXt [43] is considered as the backbone and the squeeze-and-excitation (SE) fusion module [44] is employed. Fig. 2 shows the network architecture. For the M-Net, image patches are processed by the first three residual convolutional blocks, and corresponding results are named as Res1, Res2, and Res3, respectively. Then three branches of multichannel features are aggregated and imported to the SE block for achieving feature fusion. Meanwhile, features of different modalities are outputted by the SE block from shallow to deep. This strategy makes our model better of finding intermodal correlations at both shallow and deep layers. Finally, all features from SE models are repeatedly concatenated, and then the concatenated feature maps are fed into a block composed of a global average pooling layer (GAP) and two fully connected layers. In addition, the structure of D-Net is similar to M-Net, where the difference is the inputting image size since the former requires the image patches while the latter requires the original depth images.

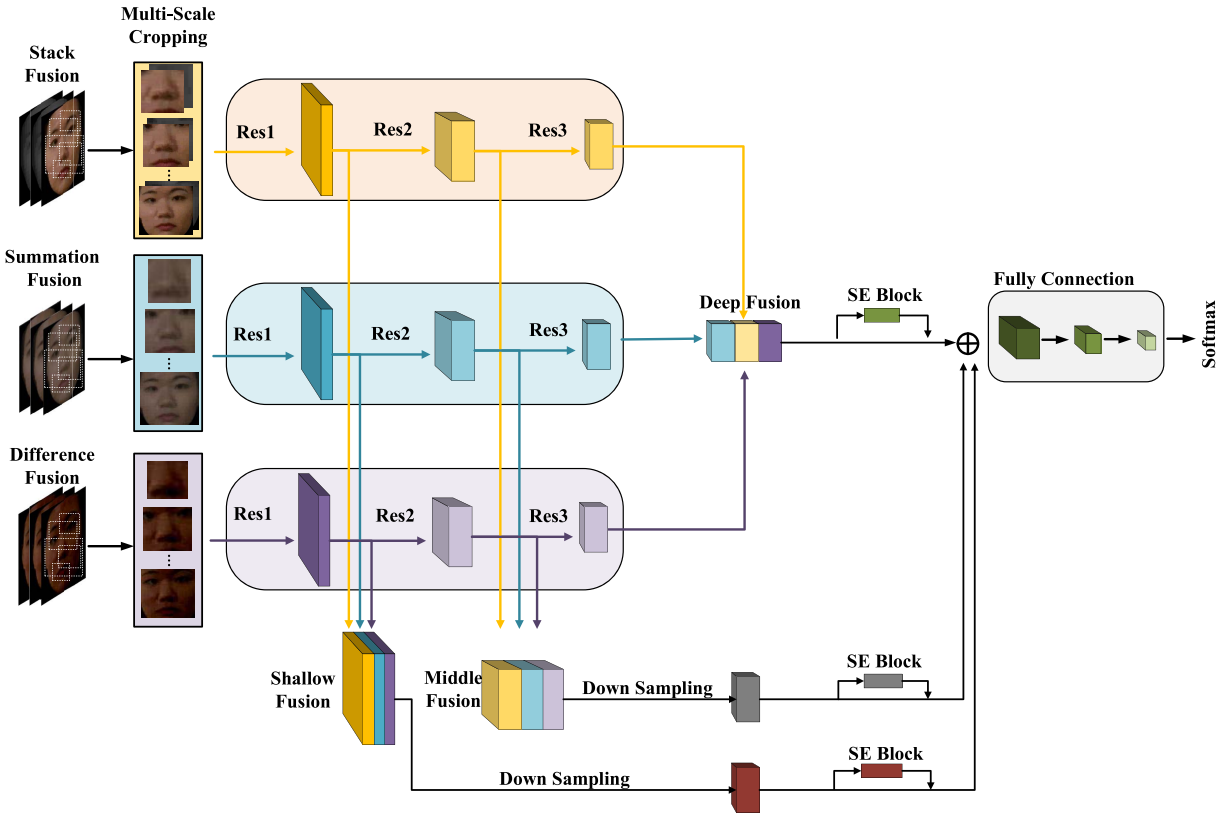


Fig. 2. Architecture of the proposed multimodality face anti-spoofing network (M-Net).

For our network design, three of data fusion modalities are jointly taken as input for PAs. To enhance the channel interdependencies among shallow, middle, and deep fusion branches, we, therefore, take the SE block to recalibrate each aggregation modality feature to exploit more useful features through subsequent transformations. The details of SE block on deep fusion branch are shown in Fig. 3.

D. Multiscale Patch and Weighted Fine-Tuning

In [45], patches with fixed size are randomly extracted from cropped face areas, which can strengthen the ability of CNNs for perceiving spoof clues on the texture attribute other than the whole facial structure attribute. However, it is time consuming to traverse all patches for network inference. To solve the problem, we propose a new fine-tuning strategy named weighted fine-tuning that can improve the discrimination ability of our network on specific patches. The weighted fine-tuning is illustrated in Algorithm 1. Specifically, we use an AdaBoost-like algorithm [46] to assign large weights to good prediction patches and small weights to poor prediction patches. Afterward, we pick out the best classifier that outputs the highest anti-spoofing response on that of selected patches in each training epoch. Then, we update all loss function weights based on the best classifier for next epoch training and update the classifier weights as well. The final prediction depends on the ensemble CNN weighted classifiers, and the input of each weighted classifier is the specific patch with predefined location. The results of the fine-tuning strategy will be explained in detail in our experiments.

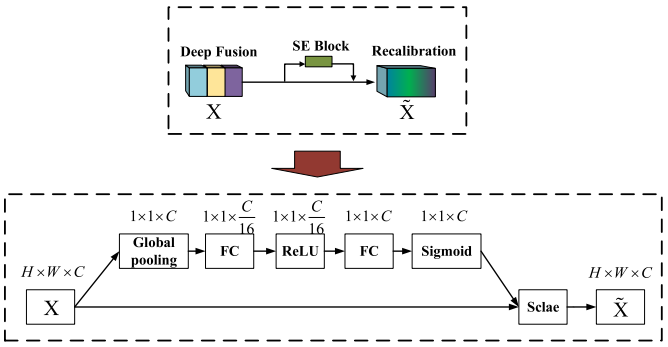


Fig. 3. Feature recalibration on deep fusion branch with SE module. *Squeeze:* A global average pooling is taken to obtain a channelwise descriptor. *Excitation:* Fully connected (FC) layers, a ReLU, and another FC layer are taken for dimension reduction and increase. The final output is obtained by rescaling with the sigmoid activations.

IV. Experiments

In this section, we conducted experiments to demonstrate the effectiveness of the proposed framework. To assess the generalization ability of each independent network, we built an MIPD. In the end, the proposed cascade network and comparative models are implemented and tested on the overall MIPD.

A. Data Sets and Metrics

For MIPD, each face includes three modalities, i.e., the RGB, depth, and IR. Figs. 4 and 5 show some samples. On the one hand, we simultaneously captured the calibrated depth,

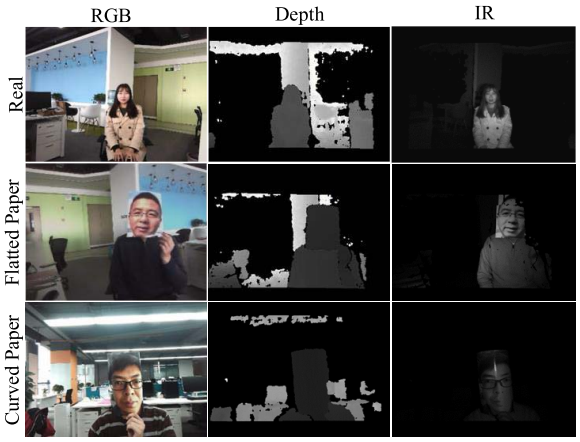


Fig. 4. Fake samples of MIP-2-D data set. The top row shows the real sample, and the middle and bottom rows show fake samples captured by flatted and curved papers.

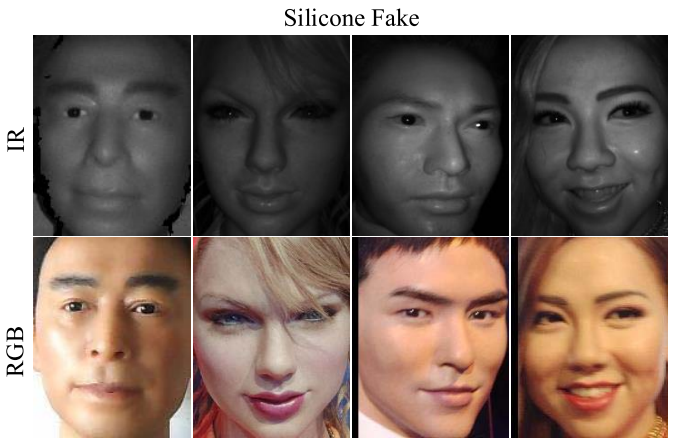


Fig. 5. Fake samples of MIP-3-D data set. The top row shows IR fake samples; and the bottom row shows RGB fake samples.

TABLE I
STATISTICAL INFORMATION OF THE MIP-2-D DATA SET

	Curved		Flatted		Simulated		Total	
	Train	Test	Train	Test	Train	Test	Train	Test
Positive	1290	320	1289	684	3901	1172	6480	2176
Negative	455	283	1134	768	3612	2227	5201	3278

color, and IR modality streams. On the other hand, we collected more than six types of fake face samples that contain basic 2-D and super-realistic 3-D fake styles. The statistics of two subdata sets (MIP-2-D and MIP-3-D) are shown in Tables I and II, respectively. For the 2-D fake style, those samples are printed with three kinds of materials and are captured with flatted and curved papers. We also collected video replaying for complementary 2-D fake style. For the 3-D fake style, user-customized and life-size figure faces made in clay with wax layers, silicone, or resin materials are captured and collected as well.

The CASIA-SURF is another largest publicly available data set used for face anti-spoofing. It consists of 1000 subjects with 21 000 videos and each sample contains three modalities (i.e., RGB, depth, and IR). For simulating the human face as real as possible, six attack ways are presented by capturing reproduction photos with hollow.

Algorithm 1 Weighted Fine-Tuning

1: **Initialisation:**
- Choose N specific patches position, denoted by P_n, where $n \in (1, \ldots, N)$
- Let the cycles of epoch T equals to the number of patches N
- Set weight for each patch position as: ω_t^n, the initialisation vallue of $\omega_t^n = \frac{1}{N}$
- Let the ground truth label vector of each patch be $\mathbf{y}_m^{P_n}$, where $\mathbf{y}^{P_n} \in (0, 1)$, $m \in (1, \ldots, M)$ is the sample of each P_n

2: **For** $t = 1, \ldots, T$

3: **Do** network inference and select the best classifier G_t with regard to patch position P_n under the minimum softmax cross entropy loss: $l_t^{P_n} = -\omega_t^{P_n} \sum_{i=1}^{M} \mathbf{y}_i^{P_n} \log(G_t)$, where $G_t = softmax(CNN(\mathbf{I}^{P_n}, \Theta))$

4: **Update** all weighted loss of each P_n with regard to G_t: $\omega_{t+1}^{P_n} = \frac{\omega_t^{P_n}}{Z_t} \times \frac{1}{M} \times \sum_{i=1}^{M} \exp(-\mathbf{y}_i^{P_n} \alpha_t G_t(P_n))$, where $\alpha_t = \frac{1}{2} \ln \frac{1-l_t^{P_n}}{l_t^{P_n}}$; $Z_t = \sum_{n=1}^{N} \omega_t^{P_n} \times \frac{1}{M} \times \sum_{i=1}^{M} \exp(-\mathbf{y}_i^{P_n} \alpha_t G_t(P_n))$

5: **Do** back propagation with regard to all patches loss function based on updated weights: $L_{t+1}^{all} = \sum_{n=1}^{N} l_{t+1}^{P_n}$, where $l_{t+1}^{P_n} = -\omega_{t+1}^{P_n} \sum_{i=1}^{M} \mathbf{y}_i^{P_n} \log(G_t)$

6: **The final strong classifier:** $C(x) = \sum_{t=1}^{T} \alpha_t G_t$

TABLE II
STATISTICAL INFORMATION OF THE MIP-3-D DATA SET

	3D Mask		3D Head		Total	
	Train	Test	Train	Test	Train	Test
Positive	1289	684	3464	1767	6043	2771
Negative	1134	768	5786	3177	8375	4228

In order to compare different networks, we choose three evaluation metrics: 1) the attack presentation classification error rate (APCER) evaluates the highest error among all samples; 2) the normal presentation classification error rate (NPCER) evaluates error of live samples; and 3) the average classification error rate (ACER) evaluates the mean of APCER and NPCER. These metrics are defined as follows:

$$\text{APCER} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{21}$$

$$\text{NPCER} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{22}$$

$$\text{ACER} = \frac{\text{APCER} + \text{NPCER}}{2} \tag{23}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

TABLE III
COMPARISON RESULTS OF DIFFERENT FUSION SCHEMES ON DEPTH MODALITY OF MIP-2-D

	Feature Protocols	ACER(%)	Accurary(%)
FaceBagNet [29]	F1	1.9	0.979
FeatherNet [47]	F1	2.6	0.976
D-Net	F1	1.2	0.986
	F1&F2	1.0	0.988
	F2&F3	1.4	0.984
	F1&F3	0.8	0.990
	F1&F2&F3	**0.5**	**0.994**

We also pay attention to the true positive rate (TPR) at some fixed false positive rate (FPR). This metric evaluation approach enables to measure how many real samples pass the anti-spoofing test while accepting no more than certain percentage of spoofing attacks.

B. Implementation Details

We use PyTorch as the deep learning framework. For both D-Net and M-Net, the backbone architecture is a modified version of ResNeXt-50. We use the stochastic gradient descent (SGD), and the weight decay and momentum are set to 0.0005 and 0.9, respectively. A cyclic cosine annealing learning rate schedule is implemented with initial learning rate 0.1. All models are trained on multiple NVIDIA Tesla V100 GPU with a batch size of 256. The whole pretraining procedure takes 200 epochs and the weighted fine-tuning procedure takes 36 epochs.

C. Intradata Set Evaluation on D-Net

In this experiment, we demonstrate the effectiveness of the depth fusion strategy based on three discriminative features on MPI-2-D. Specifically, three types of depth preprocessing schemes, including the depth face normalization (F1), depth face scale embedding (F2), and normal orientation embedding (F3), are compared with multimodalities data fusion. The overall comparison results in terms of ACER and accuracy are shown in Table III. For the basic F1 depth feature, our D-Net model achieves the ACER of 1.2% and the accuracy of 98.6%, which slightly outperforms state-of-the-art methods, such as FacebagNet and FeatherNet. Furthermore, after combining F1, F2, and F3 features, the ACER reduces to 0.8% and the accuracy achieves to 99.4%, indicating the best discriminative power for detecting depth-based fake faces. Moreover, feature combination schemes, such as F1 and F2, F2 and F3, and F1 and F3, show higher performances than that of the single feature models, which explains the effectiveness of the proposed strategy of depth fusion features.

D. Intradata Set Evaluation on M-Net

For intradata set evaluation, M-Net is trained on MIP-3-D and CASIA-SURF, respectively. Comparison experiments are shown in Table IV. We can see that under the proper illumination, the RGB model shows better performance than the IR model, since RGB data contain richer texture information.

TABLE IV
INTRADATA SET EVALUATION: INDIVIDUAL MODALITIES ON MIP-3-D

Method	TPR(%) @FPR=10^{-2}	TPR(%) @FPR=10^{-3}	TPR(%) @FPR=10^{-4}	APCER(%)	NPCER(%)	ACER(%)
RGB	1	99.034	97.63	0.1802	0.8130	0.4966
IR	99.18	81.59	50.70	1.0811	0.8130	0.9740
Fusion	1	99.42	98.96	0.9100	0.0963	0.0857
Aug Fusion	**1**	**99.92**	**99.70**	**0.0903**	**0.0738**	**0.0820**

TABLE V
INTRADATA SET EVALUATION: DIFFERENT APPROACHES ON MIP-3-D

Method	FP	FN	APCER (%)	NPCER (%)	ACER (%)
Patch-based CNN [40]	50	36	3.584	4.512	2.656
ML-LPQ [49]	15	551	37.712	1.444	19.578
LiveNet [37]	140	80	12.635	5.904	9.269
FaceBagNet [29]	12	11	1.081	0.813	0.947
Ours+	**1**	**1**	**0.090**	**0.073**	**0.082**

TABLE VI
INTRADATA SET EVALUATION: DIFFERENT APPROACHES ON CASIA-SURF

Method	Modal	FP	FN	TPR(%) @FPR=10^{-2}	TPR(%) @FPR=10^{-3}	ACER(%)
ResNet-18[27]	RGB&IR	119	811	79.10	50.90	14.40
FeatherNet	RGB&IR	225	**0**	99.03	97.56	1.700
FaceBagNet	RGB&IR	166	12	98.29	85.63	1.455
Ours	RGB&IR	142	4	99.53	93.58	1.114
Ours+	RGB&IR	**33**	5	**99.86**	**98.67**	**0.296**

Among three modalities, the fusion model achieves the best performance of 0.082% (ACER), TPR=99.7% under the condition of FPR=10e-4, which manifests the effectiveness of our proposed fusion strategy.

Table V shows results of four comparative models and the proposed model on MPI-3-D. Since patch-based CNN, ML-LPQ [47], and LiveNet [34] only consider RGB images as the network input, they suffer from inferior performance with ACER up to 2.656%, 19.578%, and 9.269%, respectively. Such poor performance is attributed to low generalization ability of the networks and single RGB modality. For fair comparison, we also consider FacebagNet model by taking IR and RGB modalities as the input streams. As a result, although FaceBagNet with multipatches attains higher performance with ACER of 0.947% than models utilizing the way of full image input, the proposed M-Net achieves the best performance with 0.082% ACER and leaves only one false positive and false negative samples.

In order to examine how the proposed nonlinear data fusion strategy contributes to the performance, we consider IR and RGB modalities as the input stream to compare the performance with other three models (FeatherNet and FacebagNet and Zhang's method [25]) on the CASIA-SURF data set. As shown in Table VI, the proposed M-Net achieves ACER of 1.114% by constructing parallel IR and RGB stream as the input of our network. By using the nonlinear data fusion strategy (stack, summation, and differencing), the ACER (Ours+) is further reduced to 0.296%, and the numbers of FP and FN samples are reduced to 33 and 5, respectively.

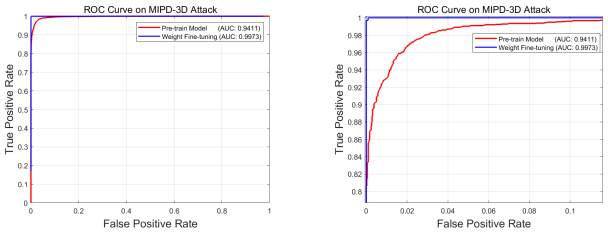


Fig. 6. ROC comparison between pretrained model and weighted fine-tuning model on MIPD-3-D data set. Left: Original ROC. Right: Zoom in ROC.

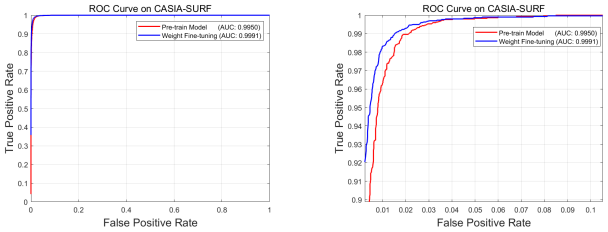


Fig. 7. ROC comparison between pretrained model and weighted fine-tuning model on CASIA-SURF data set. Left: original ROC. Right: zoom in ROC.

TABLE VII
ABLATION STUDY OF SE-BASED BACKBONE AND ASSOCIATED STRATEGIES

Backbone	Multi Scale Patch	Weighted Fine Tuning	ACER(%)
Res			0.095
Res		✓	0.091
Res	✓		0.092
Res	✓	✓	0.090
SE/Res			0.087
SE/Res		✓	0.085
SE/Res	✓		0.083
SE/Res	✓	✓	0.082

E. Weighted Fine-Tuning on M-Net

Due to the restriction of spoofing data collection, we perform weighted fine-tuning on original pretrained data sets and specific 36 patch positions for face cropping during network inference. Figs. 6 and 7 show the receiver operating characteristic (ROC) curve comparisons between the pretrained model and proposed fine-tuning model on two data sets of MIPD-3-D and CASIA-SURF. It can be shown that the weighted fine-tuning is useful due to better ROC on MIPD-3-D. Moreover, the strategy also works on CASIA-SURF since the blue curve from the weighted fine-tuning model achieves slightly higher TPR with lower FPR than the red curve from the pretrained model.

F. Ablation Study on M-Net

In this section, ablation studies are conducted on MIP-3-D to demonstrate the compelling results generated by the SE-based backbone as well as the contribution of each associated strategy. Specifically, ResNeXt-based backbone with both multiscale image patch strategy and weighted fine-tuning strategy is integrated and separately evaluated. It can be seen from the last row of Table VII that ResNeXt backbone with multilevel feature fusion outperforms (0.013% ACER) the independent ResNeXt, indicating the effectiveness of our feature skip-connection architecture. Meanwhile, it is interesting to find that the performance of our proposed network is slightly improved (0.0035% ACER on Res and 0.0030% ACER on SE/Res, respectively) by adding the multiscale patch and weighted fine-tuning strategy. From the above results, our proposed two strategies that are insensitive to the SE fusion module, hence can be promoted and further used as other networks backbone for face anti-spoofing tasks.

TABLE VIII
CROSS-EVALUATION ON D-NET BETWEEN CASIA-SURF AND MIP-2-D

Method	Train	Test	Train	Test
	CASIA-SURF	MIP-2D	MPI-2D	CASIA-SURF
ML-LPQ	56.3%		48.9%	
Patch-based CNN	45.8%		25.8%	
LiveNet	46.6%		21.1%	
FacebagNet	**43.1%**		20.2%	
Ours+	43.5%		**19.4%**	

TABLE IX
CROSS-EVALUATION ON M-NET BETWEEN CASIA-SURF AND MIP-3-D

Method	Train	Test	Train	Test
	CASIA-SURF	MIP-3D	MPI-3D	CASIA-SURF
ML-LPQ	56.9%		49.4%	
Patch-based CNN	18.7%		18.3%	
LiveNet	13.8%		12.6%	
FacebagNet	13.2%		11.5%	
Ours+	**12.6%**		**9.5%**	

G. Cross-Evaluation on CASIA-SURF and MIP-2-D/3-D

Although cross-testing is known to be substantially harder than intratesting, we still utilize CASIA-SURF, MIP-2-D, and MIP-3-D to perform cross-testing. Tables VIII and IX show cross-testing comparison results of different models for the D-Net and M-Net, respectively. From Tables VIII and IX, we can draw the following conclusions.

1) The overall ACER scores in Table VIII are higher than in Table IX, indicating that the model generalization ability on D-Net is lower than M-Net.
2) In the case of cross-testing between CASIA-SURF and MIP-2-D, the D-Net model does not perform well since the external background of face images on CASIA-SURF is all replaced by uniform zero values, whereas no further images preprocessing is implemented on MIP-2-D. Hence, the different ways of images preprocessing finally impact on ACER by dramatically increasing false positive samples.
3) For the M-Net model comparison shown in Table IX, the proposed model achieves the lowest ACER of 12.6% and 9.5% when the flipping data set is used for training and testing. Besides, the cross-data set evaluation did not perform as good as intradata set evaluation due to the diversity of the image quality, such as noise, resolution, exposure, etc.

H. Cross-Evaluation on M-Net Between MIP-2-D and MIP-3-D

To evaluate model impact by different types of fake style variations, we perform cross-data set evaluation on MIP-2-D

TABLE X
CROSS-EVALUATION ON M-NET WITH MIP-3-D TRAINING AND MIPD TESTING

Method	FP	FN	APCER (%)	NPCER (%)	ACER (%)
ResNet-18[27]	64	57	0.858	1.115	0.983
FeatherNet	**26**	34	**0.346**	0.687	0.517
FaceBagNet	63	10	0.839	0.202	0.521
CDCN [16]	12	11	1.081	0.813	0.947
Ours+	45	**6**	0.599	**0.121**	**0.301**

TABLE XI
COMPARISON RESULTS OF CASCADE NETWORK AND OTHER METHODS ON OVERALL MIPD DATA SET

Method	FP	FN	APCER (%)	NPCER (%)	ACER (%)
ResNet-18[27]	32	20	0.4263	0.4042	0.4152
FeatherNet	15	24	0.1998	0.4851	0.3425
FaceBagNet	12	**3**	0.1598	**0.0606**	0.1102
Ours+	**8**	6	**0.0932**	0.1211	**0.1071**

and MIP-3-D. To follow our original intention to establish a cascade network that can resist 2-D and 3-D-based face anti-spoofing, we evaluate the M-Net by the training model on MIP-3-D and testing model on MIPD. As shown in Table X, our proposed M-Net with data fusion strategy (Ours+) achieves the ACER of 0.301, which outperforms comparative models, including CDCN, FaceBagNet, FeatherNet, and ResNet-18. To avoid the uncertainty of illumination from RGB images, we choose to take input of the IR image for CDCN. From the results on our model, we can see that the FP number raised greater than the FN number compared to intradata set evaluation in Table V. It further manifests that fake faces with different materials will affect the performance of M-Net. Hence, it demonstrates the necessity of designing a cascade network to eliminate different types of spoofing faces.

I. Cascade Evaluation

For the cascade framework evaluation, each hierarchical network is individually trained on the MIP-2-D and MIP-3-D, respectively. The above experimental results show that both D-Net and M-Net only provide limited performance for face anti-spoofing. In fact, the D-Net can be viewed as a preliminarily discriminator to prevent basic face spoofing styles, such as flatten or curved papers, videos, etc., and M-Net plays a role of decisive discriminator for final classification based on the intrinsic material attribute. It is clear that the proposed cascade framework integrates advantages of D-Net and M-Net. Table XI shows that the proposed cascade framework with ACER of 0.1071% outperforms all three comparative models with ACER of 0.4152%, 0.3425%, and 0.1102%, respectively.

V. DISCUSSION

With multimodality streams, conventional printed-like 2-D face attacks can be easily discriminated by D-Net based on depth data, and vivid 3-D face attacks can be also prevented subsequently by M-Net based on multimodality data. Due to the fact that many factors comprehensively affect the face anti-spoofing system, we did serval functionally independent subexperiments to assess the contribution of different factors. For our proposed cascade framework, the first stage of D-Net only focuses on the depth stream. Hence, it ensures most of 2-D spoofing faces to be perfectly recognized and merely allows 3-D realistic fake faces to be identified at the next M-Net. Due to such framework design, the collection of training data for M-Net is composed of large sets of 3-D realistic fake samples and small sets of 2-D fake samples, which improves the discrimination ability of M-Net on 3-D fake samples but not fails to prevent 2-D fake samples. From the experimental results, it can been seen that M-Net presents better results compared to state-of-the-art models due to better nonlinearity from multimodality data and better feature fusion from multiple network levels. Furthermore, the weighted fine-tuning strategy improves network performance because we do not bring any extra training samples but merely take specific patches from original images as fine-tuning data. Actually, the strategy can be also adopted by taking our network as a pretrained model in other face anti-spoofing tasks.

VI. CONCLUSION AND FUTURE WORK

In this work, we have studied the task of face anti-spoofing for preventing both 2-D and 3-D face attacks under several identification verification scenarios. For this task, we have developed a two-stage cascade framework to extract both face reflectance features and multilevel of face texture features by considering the data nonlinearity fusion strategy and network skip-connection architecture. The experimental results show that the proposed anti-spoofing framework can prevent diversity of face attacking forms, such as dim light, realistic face camouflage, static or motion pattern, etc. Furthermore, the proposed model shows strong generalization ability on PAs since it fuses features from coarse to fine network levels and utilizes the nonlinearity of multimodality information. For future works, we will establish a more pervasive face spoofing data set to analyze the generalization ability of the proposed framework. Moreover, the proposed cascade strategy can also be extended toward other tasks of biometric modality attack detection, such as print attack in iris and palm.

REFERENCES

[1] I. Chingovska, A. R. dos Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2264–2276, Dec. 2014.
[2] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 20–30, Sep. 2015.
[3] X. Zhu, S. Li, X. Zhang, H. Li, and A. C. Kot, "Detection of spoofing medium contours for face anti-spoofing," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 28, 2020, doi: 10.1109/TCSVT.2019.2949868.
[4] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 7, pp. 1713–1723, Jul. 2017.
[5] 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.
[6] J. Galbally and S. Marcel, "Face anti-spoofing based on general image quality assessment," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Stockholm, Sweden, 2014, pp. 1173–1178.

[7] 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.

[8] P. P. K. Cha *et al.*, "Face liveness detection using a flash against 2D spoofing attack," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 521–534, Feb. 2018.

[9] A. Z. Abd Aziz, H. Wei, and J. M. Ferryman, "Face anti-spoofing countermeasure: Efficient 2D materials classification using polarization imaging," in *Proc. Int. Workshop Biometr. Forensics (IWBF)*, Coventry, U.K., 2017, pp. 1–6.

[10] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *Proc. Int. Conf. Biometr. (ICB)*, Madrid, Spain, 2013, pp. 1–6.

[11] J. Gan *et al.*, "Spatial-temporal texture cascaded feature method for face liveness detection," *Pattern Recognit. Lett.*, vol. 32, no. 2, pp. 117–123, 2019.

[12] 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.

[13] T. A. Siddiqui *et al.*, "Face anti-spoofing with multifeature videolet aggregation," in *Proc. 25rd Int. Conf. Pattern Recognit. (ICPR)*, 2017, pp. 1035–1040.

[14] X. Yang *et al.*, "Face anti-spoofing: Model matters, so does data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 3507–3516.

[15] 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 Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, 2018, pp. 544–552.

[16] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection using 3D structure recovered from a single camera," in *Proc. Int. Conf. Biometr. (ICB)*, Madrid, Spain, 2013, pp. 1–6.

[17] Z. Yu *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* Seattle, WA, USA, 2020, pp. 5294–5304.

[18] F. Jiang, P. Liu, and X. Zhou, "Multilevel fusing paired visible light and near-infrared spectral images for face anti-spoofing," *Pattern Recognit. Lett.*, vol. 128, pp. 30–37, Dec. 2019.

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

[20] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4726–4740, Dec. 2015.

[21] G. Heusch, T. de Freitas Pereira, and S. Marcel, "A comprehensive experimental and reproducible study on selfie biometrics in multistream and heterogeneous settings,"*IEEE Trans. Biometr. Behav. Identity Sci.*, vol. 1, no. 4, pp. 210–222, Oct. 2019.

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

[23] A. J. Liu *et al.*, "Cross-ethnicity face anti-spoofing recognition challenge: A review," 2020. [Online]. Available: arXiv: 2003.10998

[24] A. Liu *et al.*, "Static and dynamic fusion for multi-modal cross-ethnicity face anti-spoofing," 2020. [Online]. Available: arXiv: 1912.02340

[25] S. Zhang *et al.*, "CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Trans. Biometr. Behav. Identity Sci.*, vol. 2, no. 2, pp. 182–193, Apr. 2020.

[26] A. Parkin and O. Grinchuk, "Recognizing multi-modal face spoofing with face recognition networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 1617–1623.

[27] T. Shen, Y. Huang, and Z. Tong, "FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 1611–1616.

[28] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu, "PipeNet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, 2020, pp. 644–645.

[29] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu and A. K. Nandi, "DefED-Net: Deformable encoder-decoder network for liver and liver tumor segmentation," *IEEE Trans. Radiat. Plasma. Med. Sci.*, early access, Feb. 16, 2020, doi: 10.1109/TRPMS.2021.3059780.

[30] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, New York, NY, USA, 2020, pp. 11974–11981.

[31] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," in *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.

[32] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," in *IEEE Trans. Circuits Syst. Video Technol.* vol. 30, no. 3, pp. 734–747, Mar. 2020.

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

[34] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 389–398.

[35] Y. A. U. Rehman, P. L. Man, and M. Liu, "LiveNet: Improving features generalization for face liveness detection using convolution neural networks," *Expert Syst. Appl.*, vol. 108, pp. 159–169, Oct. 2018.

[36] A. Alotaibi and A. Mahmood, "Deep face liveness detection based on nonlinear diffusion using convolution neural network," *Signal Image Video Process.*, vol. 11, no. 4, pp. 713–720, 2017.

[37] L. Feng *et al.*, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *J. Vis. Commun. Image. Represent.*, vol. 38, pp. 451–460, Jul. 2016.

[38] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. Int. Joint Conf. Biometr. (IJCB)*, Denver, CO, USA, 2017, pp. 319–328.

[39] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 4885–4894.

[40] D. P. Playne and K. Hawick, "A new algorithm for parallel connected-component labelling on GPUs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 6, pp. 1217–1230, Jun. 2018.

[41] S. Alotaibi and W. A. P. Smith, "Decomposing multispectral face images into diffuse and specular shading and biophysical parameters," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, 2019, pp. 3138–3142.

[42] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prjgel-Bennett, and J. Hare, "Understanding and enhancing mixed sample data augmentation," 2020. [Online]. Available: arXiv: 2002.12047

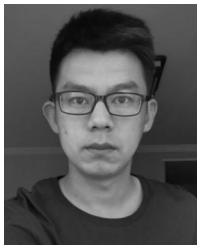
[43] S. Xie, R. B. Girshick, P. Dollár,, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 5987–5995.

[44] X. Zhong, O. Gong, W. Huang, L. Li, and H. Xia, "Squeeze-and-excitation wide residual networks in image classification," in *Proc. Int. Conf. Image Process. (ICIP)* Taipei, Taiwan, 2019, pp. 395–399

[45] P. Zhang *et al.*, "FeatherNets: convolutional neural networks as light as feather for face anti-spoofing," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 1574–1583.

[46] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[47] A. Benlamoudi, D. Samai, A. Ouafi, S. E. Bekhouche, A. Taleb-ahmed, and A. Hadid, "Face spoofing detection using multi-level local phase quantization (ML-LPQ)," in *Proc. Int. Conf. Autom. Control Telecommun signals (ICATS)*, Annaba, Algeria, 2015, pp. 1–5.



Weihua Liu received the Ph.D. degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2014.

From 2011 to 2013, he was a Visiting Scholar with the Vision-Learning-Mining Research Laboratory, University of Texas at Arlington, Arlington, TX, USA. From 2014 to 2016, he was an Assistant Researcher with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Beijing, China. From 2016-2020, he was a Research Fellow with the Laboratory of Intelligent Image Processing, Orbbec Company, Shenzhen, China. He is currently an associate professor with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, and also a Research Fellow with the Laboratory of Intelligent Image Processing, Orbbec Company. His research interests include image processing, machine learning, and computer vision.

Xiaokang Wei received the bachelor's degree in process equipment and control engineering from the Dalian University of Technology, Dalian, China, in 2017, and the master's degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2020.

He is currently a Research Fellow in the Innovation Lab, Orbbec, Shenzhen, China. From 2018 to 2019, he was a Research Intern with Orbbec Innovation Lab, Troy, MI, USA. He has coauthored a research paper in AIFT and has published two Chinese National Invention Patent. His research interests include 3-D vision, graphics, and deep learning.

Tao Lei (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2011.

From 2012 to 2014, he was a Postdoctoral Research Fellow with the School of Electronics and Information, Northwestern Polytechnical University. From 2015 to 2016, he was a Visiting Scholar with the Quantum Computation and Intelligent Systems Group, University of Technology Sydney, Sydney, NSW, Australia. He is currently a Professor with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an. He has authored and coauthored over 90 research papers, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON RADIATION AND PLASMA MEDICAL SCIENCES, ICASSP, ICIP, and FG. His current research interests include image processing, pattern recognition, and machine learning.

Xingwu Wang received the bachelor's degree in information security from Xidian University, Xi'an, China, in 2019. He is currently pursuing the M.S. degree with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an.

His current research interests include image processing and pattern recognition.

Hongying Meng (Senior Member, IEEE) received the Ph.D. degree in communication and electronic systems from Xi'an Jiaotong University, Xi'an, China, in 1998.

He is currently a Reader with the Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge, U.K. He has authored over 130 publications, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON AUTOMATIC CONTRO, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, ICASSP, and NIPS. His research interests include digital signal processing, affective computing, machine learning, human–computer interaction, and computer vision.

Dr. Meng is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEOS TECHNOLOGY and IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS. He is a Fellow of The Higher Education Academy and a member of Engineering Professors Council, U.K.

Asoke K. Nandi (Fellow, IEEE) received the Ph.D. degree in physics from the Trinity College, University of Cambridge, Cambridge, U.K., in 1978.

He held academic positions in several universities, including the University of Oxford, Oxford, U.K.; Imperial College London, London, U.K.; the University of Strathclyde, Glasgow, U.K.; and Liverpool John Moores University, Liverpool, U.K., as well as Finland Distinguished Professorship in Jyväskylä University, Jyväskylä, Finland. In 2013, he moved to Brunel University London, Uxbridge, U.K., to become the Chair and the Head of Electronic and Electrical Engineering. He is a Distinguished Visiting Professor with Xi'an Jiaotong University, Xi'an, China, and an Adjunct Professor with the University of Calgary, Calgary, AB, Canada. He has authored over 600 technical publications, including 250 journal papers as well as five books, titled *Condition Monitoring With Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines* (Wiley, 2020), *Automatic Modulation Classification: Principles, Algorithms and Applications* (Wiley, 2015), *Integrative Cluster Analysis in Bioinformatics* (Wiley, 2015), *Blind Estimation Using Higher-Order Statistics* (Springer, 1999), and *Automatic Modulation Recognition of Communications Signals* (Springer, 1996). The H-index of his publications is 80 (Google Scholar) and his ERDOS number is 2. He has made many fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "Big and Heterogeneous Data," dealing with modeling, classification, estimation, and prediction. His current research interests lie in signal processing and machine learning, with applications to communications, image segmentations, and biomedical data.

Dr. Nandi received the Institute of Electrical and Electronics Engineers (USA) Heinrich Hertz Award in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers (U.K.) in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (U.K.) in 1998. In 1983, he co-discovered the three fundamental particles known as W^+, W^-, and Z^0 (by the UA1 team at CERN), providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics in 1984 Awarded the prize to his two team leaders for their decisive contributions. He is an IEEE EMBS Distinguished Lecturer from 2018 to 2019. He is a Fellow of the Royal Academy of Engineering (U.K.) as well as a Fellow of seven other institution, including IET.