Taming Self-Supervised Learning for Presentation Attack Detection: De-Folding and De-Mixing

Zhe Kong, Wentian Zhang, Feng Liu, and Linlin Shen are with the College of Computer Science and Software Engineering and the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518107, China (e-mail: whluo.china@gmail.com).

Haozhe Liu was with the Norwegian University of Science and Technology (NTNU), 2818 Gjøvik, Norway. He is now with the College of Computer Science and Software Engineering and the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, SZU Branch, Shenzhen 518060, China (e-mail: whluo.china@gmail.com).

Raghavendra Ramachandra is with the Norwegian Biometrics Laboratory (NBL), Norwegian University of Science and Technology, 2818 Gjøvik, Norway (e-mail: raghavendra.ramachandra@ntnu.no).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2023.3243229, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2023.3243229

Abstract—Biometric systems are vulnerable to presentation attacks (PAs) performed using various PA instruments (PAIs). Even though there are numerous PA detection (PAD) techniques based on both deep learning and hand-crafted features, the generalization of PAD for unknown PAI is still a challenging problem. In this work, we empirically prove that the initialization of the PAD model is a crucial factor for generalization, which is rarely discussed in the community. Based on such observation, we proposed a self-supervised learning-based method, denoted as DF-DM. Specifically, DF-DM is based on a global–local view coupled with de-folding and de-mixing to derive the task-specific representation for PAD. During de-folding, the proposed technique will learn region-specific features to represent samples in a local pattern by explicitly minimizing the generative loss. While de-mixing drives detectors to obtain the instance-specific features with global information for more comprehensive representation by minimizing the interpolation-based consistency. Extensive experimental results show that the proposed method can achieve significant improvements in terms of both face and fingerprint PAD in more complicated and hybrid datasets when compared with the state-of-the-art methods. When training in CASIA-FASD and Idiap Replay-Attack, the proposed method can achieve an 18.60% equal error rate (EER) in OULU-NPU and MSU-MFSD, exceeding the baseline performance by 9.54%. The source code of the proposed technique is available at https://github.com/kongzhecn/dfdm.

Index Terms—Presentation attack detection (PAD), self-supervised learning.

I. INTRODUCTION

With the applications in mobile phone unlocking, access control, payment tool, and other security scenarios, biometric systems are widely used in our daily lives. Among the most popular biometric modalities, fingerprint and face play vital roles in numerous access control applications. However, several reported studies [1], [2] have demonstrated that the existing systems are easily spoofed by presentation attacks (PAs) made from low-cost materials and instruments, e.g., Rigid Mask for face [3] and silica gel for fingerprint [4]. These issues raise wide concerns about the vulnerability of biometric systems incorporated in access control applications. Therefore, it is essential to detect PAs to achieve reliable biometric applications.

To reliably address the vulnerability of the biometric systems to PAs, several PA detection (PAD) methods have been proposed [1], which can be divided into hardware- and software-based methods. The hardware-based solutions [3], [5], [6] use special types of sensors to capture liveness characteristics. For instance, Heusch et al. [3] adopt short-wave infrared (SWIR) imaging technology to detect face PAs, which shows superior performance over similar models working on color images. A light field camera (LFC) is introduced by Raghavendra et al. [6] to detect PAs by exploring the variation in the focus between multiple depth images. For fingerprints, an optical coherence tomography (OCT)-based PAD system is designed by Liu et al. [5] to obtain the depth information of fingerprints. Generally speaking, hardware-based solutions are sensor-specific, resulting in strong security but weak applicability because of usability or cost limitations, and the current mainstream is the software-based methods.

Fig. 1 illustrates the recent progress on the software-based PAD algorithms that can be categorized into three groups: 1) input preprocessing; 2) model design; and 3) loss function. In the case of input preprocessing [7], [8], [9], [10], [11], Larbi et al. [8] propose a model, namely, DeepColor-FASD model, which adopts various color spaces (RGB, HSV, YCbCr) as input to achieve the reliable performance of PAD. Despite the improvement, the additional color spaces need...
Fig. 1. Groups of the software-based PAD. Group 1 is input preprocessing, Group 2 refers to model designing, and Group 3 is loss function designing. Different from the existing groups, the proposed method investigates the initialization of the PA detector, which can be concluded as an independent solution, i.e., Group 4: parameter initializing.

| TABLE I | AVERAGE PERFORMANCE OF PA DETECTOR WITH DIFFERENT INITIALIZATION IN TERMS OF EER (%) ↓, AUC (%) ↑, AND TDR(%)@FDR = 1.0% ↑ UNDER THE CROSS-DATASET SETTING OF THE FACE. MORE DETAILS CAN BE FOUND IN THE APPENDIX |
|---------|-----------------------------------------------------------------------------------|
| Group1: | OULU-NPU[7], MSU-MFSD[8]                                                                 |
| Metrics | Group2: CASIA-PASS[9]                                                              |
| EER (%) | AUC (%) | TDR (%) |
| Trained from Scratch | 42.3  | 73.51  | 4.41  |
| Pre-trained from ImageNet | 31.28 | 73.68  | 10.06 |
| DF-DM  | 18.78  | 89.62  | 30.39  |

For these software-based PAD algorithms, an important interference factor is the initialization of the PA detector. Generally, training from scratch and pretraining using ImageNet are two common methods. In terms of PAD, it is challenging to collect large-scale data; hence, without any prior knowledge, it is difficult to train the model from scratch (i.e., random initialization) to learn discriminative features. As listed in Table I, without any pretraining strategies, the detector can only reach 42.30% of equal error rate (EER) for face anti-spoofing. While the detector pretrained from ImageNet can achieve 31.28% EER, performing higher generalization against different datasets. Such empirical results indicate that initialization plays a vital role in improving the generalization of PAD. However, taking a pretrained model from ImageNet as initialization is also not a proper choice. As a large-scale dataset, the cost of time and computation carried on ImageNet is an over-heavy workload to train new proposed PAD CNN architectures. Meanwhile, face and fingerprint images are quite different from natural images in both texture and context, and the pretrained model is thus not an ideal and reasonable starting point for the PAD task.

To solve the addressed problems, a self-supervised learning method, denoted as DF-DM, is proposed in this article. Without any PAD labels, two pretext tasks are designed to train the network for the initialization of the PA detector. Based on the chirality of fingerprints and the symmetry of faces, a generative task denoted as de-folding is designed to force a CNN-based model to reconstruct the folded images by learning the specific patterns among various regions in faces or fingerprints. To facilitate the extraction of global features, another pretext task de-mixing is proposed.

In the de-folding task, the texture of fingerprints or faces is all overlapped in the folded images. The network is trained to dissociate the chaotic texture in the folded images. In this process, the network learns texture-related semantics, which is helpful for the PAD tasks. Due to the chirality of fingerprint images, a fingerprint image is folded in the horizontal and vertical directions, respectively. But the face is left–right symmetrical, and the face images are only folded once in the vertical direction to learn the symmetrical features of the face. In the de-folding task, the model learns to represent the
images with region-specific features in a local view. To further strengthen the representation between samples, the de-mixing task is proposed to facilitate the extraction of local features. In the de-mixing task, the network disentangles the mixed images from two samples, which learns the texture-related semantic information in a global view. The pretext tasks de-folding and de-mixing reconstruct images from local and global views, respectively. The detector can achieve 18.78% EER after using our proposed DF-DM pretext task, which gets a much better result than the model trained from scratch and pretrained from ImageNet. The proposed method, which performs in an unsupervised manner with limited computational resources, achieves impressive performance in terms of face and fingerprint PAD.

A. Fingerprint PAD Methods

For fingerprint PAD methods, the CNN-based networks can achieve satisfactory performance [4], [9], [20], [21], [22], [23]. Nogueira et al. [20] pretrain deep CNNs for object recognition and then fine-tune the CNN for fingerprint PAD. Chugh et al. [21] extract local patches centered and align the input using minutiae for the fingerprint PAD model. To improve the generalization, researchers have further proposed numerous methods to improve the performance across “unknown” or novel attacks. Pereira et al. [22] address the generalization problem by applying a regularization technique based on adversarial training. A generative adversarial network (GAN)-based data augmentation, called universal material generator (UMG), is proposed by Chugh and Jain [9] to transfer the style (texture) characteristics between fingerprint images to train a robust PA detector. Liu et al. [4] propose a global–local model-based PAD (RTK-PAD) method to overcome the information loss and improve the generalization ability.

The one-class-based approach is also proposed to address unknown attack problems. Sequeira and Cardoso [24] compare the performance of the supervised and semi-supervised approaches that rely solely on bona fide samples. Liu et al. [5] propose a one-class PAD model OCPAD, which is based on an autoencoder network. The proposed OCPAD model is learned from the training set containing only bona fide samples. The reconstruction error and latent code obtained by the trained autoencoder network are used to calculate the spoof score.

B. Face PAD Methods

The software-based face PAD methods can be categorized into hand-craft-based methods and deep-learning-based methods. LBP [13], HoG [14], SIFT [15], and Surf [16] are often...

II. RELATED WORKS

Since this article mainly focuses on PAD solutions based, we review not only the most representative works of self-supervised learning in vision tasks but also other fingerprint and face PAD methods in this section.

1) A generative task, de-folding, is designed for fingerprints and faces PAD to explore the specific patterns and textures among different regions.

2) As a complementary task, the de-mixing task with the interpolation-based loss is proposed to learn more global features to better represent images.

3) The proposed method, which performs in an unsupervised manner with limited computational resources, achieves impressive performance in terms of face and fingerprint PAD.
used to extract hand-craft features, and then these extracted features are input to a traditional classifier such as LDA and SVM for classification. Hand-crafted features are sensitive to noise, so they cannot generalize well to different illumination or different attack types.

With the success of CNN, numerous deep-learning-based face PAD methods have been proposed [7], [18], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]. Liu et al. [7] propose a CNN-RNN model to learn auxiliary features, including depth and rPPG, for PAD. Yang et al. [40] proposed a data collection method using a data synthesis technique to generate spoof samples. A deep tree network (DTN) is proposed by Liu et al. [36] to partition the spoof samples into semantic subgroups and detect PAs by routing test samples into similar clusters. Yu et al. [38] propose a central difference convolution (CDC) layer to capture intrinsic detailed patterns via aggregating both intensity and gradient information. By adopting neural architecture search (NAS), the CDC-based network can achieve superior performance. The one-class loss proposed by George and Marcel [37] tries to learn a compact embedding space for the bona fide samples. However, the one-class loss only considers the domain-invariant features and ignores the differences among domains. Hence, Jia et al. [18] propose an asymmetric triplet loss to mine the PAD features and design a single-side adversarial loss to align the mined features between the different domains. Besides, Wang et al. [32] proposed a novel shuffled style assembly network (SSAN) to extract and reassemble different content and style features for a stylized feature space. More recently, Zhang et al. [41] propose to extract the prior knowledge from the face-related works in a face system to improve the generalization of face PAD.

C. Self-Supervised Learning Methods for Vision Tasks

Self-supervised learning refers to learning methods in which CNNs are trained with automatically generated labels and then transferred to other computer vision tasks [42], [43], [44], [45]. Based on the categories of the generated labels, self-supervised learning can be roughly divided into generative and contrastive learning [46]. However, both kinds of methods cannot be directly used in PAD. As fingerprint and face images for recognition lack colorful information, and generally with low resolution, generative learning, such as image colorization [47] and image super-resolution [48], cannot be conducted in this case. Meanwhile, for image in-painting [49], and GANs [50], [51], large-scale data are required to establish a compact feature space, while the dataset for PAD cannot meet the requirement. On the other side, face images have strong spatial specifications after alignment, while the spatial relationship of fingerprints is weak. Hence, contrastive learning, like predicting the relative position [52] and rotation [53], is easy for the face but too hard for fingerprint. Another group of contrastive learning is instance discrimination, like MoCo [54], [55], SimCLR [56], and BYOL [57]. Through embedding each instance/image into different classes, the mentioned studies have shown solid improvement in natural images. However, in the PAD task, the bijective relationship between each image and prediction leads the model to learn identification rather than PAD features, resulting in poor generalization performance [58].

The self-supervised learning methods for PAD have been proposed [59], [60], [61]. Wang et al. [59] reformulate face anti-spoofing as a fine-grained patch-type recognition task and present a simple training framework called PatchNet to efficiently learn the embedding of the patch with the spoof-related capture characteristics. Wang et al. [59] and our proposed method all emphasize that texture or structural materials play an important role in PAD. But PatchNet adopts a recognition task to learn these texture or materials features, while we propose a reconstruction task (de-folding) to learn the texture and material features. Furthermore, Wang et al. [61] propose a novel embedding-level and prediction-level consistency regularization method for deep face anti-spoofing. The consistency of the two feature maps, extracted from the same input but with different augmentation, is then used to boost the PAD model. But in our de-mixing tasks, we compute the distance of two feature maps to learn the relationship between the samples, and the two feature maps in the de-mixing task are derived from two different input images. When the training data are videos, Muhammad et al. [60] propose temporal sequence sampling (TSS) for 2-D face PAD by removing the estimated interframe 2-D affine motion in the view and encoding the appearance and dynamics of the resulting smoothed video sequence into a single RGB image.

In this article, a novel self-supervised learning, namely, DF-DM, is proposed to improve the performance of PAD. Unlike the existing PAD methods, the proposed method is free of any PAD labels and extra data and pays more attention to the specification of the face and fingerprint. Two pretext tasks, de-mixing and de-folding, are proposed to search for a reasonable initialization for the PA detector. The de-folding task explores the properties of the face and fingerprint, such as chirality and symmetry, by searching the differences among the patches from a given sample. In contrast, de-mixing requires the model to embed the samples into a compact but distinguishable feature space by localizing the relationship between the different samples. By drawing de-folding and de-mixing simultaneously, adequate PAD features are extracted, which can be useful for detecting PAs. Extensive experiments clearly show significant improvements in the performance of face and fingerprint PAD.

III. PROPOSED METHOD

Fig. 2 presents the block diagram of the proposed DF-DM, which adopts de-folding and de-mixing to reliably capture the hierarchical features useful for PAD. The goal of de-folding is to reconstruct the raw image from the folded image. Since the folded image and the corresponding ground truth can be easily obtained, the model in this task is directly trained by minimizing the generative losses in an explicit way. While the de-mixing task is an ill-posed problem, where a single mixed image corresponds to two different images (irrespective of the order). Hence, we introduce a new loss function called
interpolation-based consistency to train the model for de-mixing in an implicit way. In Sections III-A–III-C, we will present a detailed discussion of the proposed method.

A. De-Folding Task: Searching Differences Among the Patches

The patterns of the face and fingerprint are quite different from that of natural images. A typical case of the point is that the fingerprint and face perform symmetric distribution in the global view of the images but chirality in the local patterns, such as texture features and reflection. In PAD, print photographs, replay videos, and 3-D masks are typical attacks for the biometric recognition systems. Although the attacks are similar to the bona fide samples from the view of human vision, the texture of the attacks is generally unusual, with anomalous reflection due to the specification of the instruments. As the features of PAD are mainly identical to the chiral features, a chirality-related pretext task, denoted as de-folding, is proposed in this article.

As shown in Fig. 3, based on different modalities, we propose two strategies to fold images. In the case of the face, a vertical line is adopted to cut the input image into two patches, $A_1$ and $A_2$, which are then randomly selected to flip horizontally to obtain $A'_1$ and $A'_2$. Through resizing and averaging $A'_1$ and $A'_2$, the folded image $f_i$ can be calculated, which is then drawn as the input of the following part. Unlike the left–right symmetrical face, the fingerprint shows chirality in the vertical and horizontal directions. Hence, $x_i$ is cropped into four patches, $\{A_1, A_2, A_3, A_4\}$, by the vertical and horizontal lines. And the flips with various directions are correspondingly adopted to generate $\{A'_1, A'_2, A'_3, A'_4\}$. To improve the difficulty of the task and prevent the model from overfitting, the lines for cutting are randomly localized rather than frozen in the middle of the image.

Since the paired data, i.e., $(f_i, x_i)$ for de-folding, can be generated easily, the model is trained explicitly by minimizing the generative losses. In particular, given $f_i$ as input, a feature extractor $D(\cdot)$ is adopted to embed $f_i$ into a latent representation $z_i$, while a generator $G(\cdot)$ is used to reconstruct $z_i$ to $y_i$. By following such cycle pipeline $x_i \rightarrow f_i \rightarrow y_i, D(\cdot)$ and $G(\cdot)$ are trained end-to-end by the learning objective:

$$
\min_{G, D} \mathcal{L}_D (y_i, x_i) + \mathcal{L}_G (y_j, x_i)
$$

where $\mathcal{L}_D (y_i, x_i) = \mathbb{E}_{x_i \sim \mathcal{X}_i} \left[ ||y_i - x_i||_2 \right]$ and $\mathcal{L}_G (y_j, x_i) = \mathbb{E}_{x_i \sim \mathcal{X}_i} \left[ F(x_i) - \mathbb{E}_{f_j \sim \mathcal{T}(x_i)} [F(y)] \right] \quad (1)$

B. De-Mixing Task: Preserving the Relationship Among Different Samples

Since de-folding performs in a local view, the model leans to represent the images with region-specific features for reconstruction. Such a pretext task pays more attention to the local patterns but neglects the relationship between the samples. As a result, varying samples can be embedded into similar representations. However, PAD is a binary classification task in which the ideal embedding space is compact but distinguishable for different samples. Therefore, in this article, another pretext task, denoted as de-mixing, is proposed to further enhance the discrimination among the different samples. Inspired by the work [65], [66], the model is required not only to reconstruct folded images but also to disentangle mixed images from different samples. In De-Mixing task, two samples $x_i$ and $x_j$ are mixed into $M_{ij}$ by

$$
M_{ij} = \Delta_{\epsilon \sim \mathbb{U}(0,1)} (x_i, x_j) = \epsilon x_i + (1-\epsilon)x_j \quad (2)
$$

where $\mathbb{U}(0,1)$ is the uniform distribution from 0 to 1., and $\epsilon$ is a scalar sampled from $\mathbb{U}(0,1)$ for mixing. Given $M_{ij}$ as input, the feature extractor $D(\cdot)$ is required to disentangle $M_{ij}$ into $x_i$ and $x_j$. However, such a requirement makes the task an ill-posed problem, which is hard to train end-to-end. Considering the ground truth of de-mixing, both $\{x_i, x_j\}$ and $\{x_j, x_i\}$ are the correct results. But for $D(\cdot)$, the order changing in the ground truth is regarded as different labels. To overcome the problem, the de-mixing task is trained in an implicit way using interpolation-based consistency $\mathcal{L}_t$

$$
\mathcal{L}_t (x_i, x_j) = \mathbb{E}_{x_i, x_j \sim \mathcal{X}_i} [||z_{ij} - \tilde{z}_{ij} + \delta||_2]
$$

$$
\tilde{z}_{ij} = \Delta(z_i, z_j) \quad (3)
$$

where $\delta$ is a random noise sampled from a Gaussian distribution with 0. mean and 0.1 standard deviations. By minimizing the distance between $z_{ij}$ and $\tilde{z}_{ij}$, the mixing operation is identical in both image and embedding space. Since $\{z_i, z_j, z_{ij}\}$ has the same topological structure as $\{x_i, x_j, M_{ij}\}$, $M_{ij}$ can be de-mixed easily in the embedding space of $D(\cdot)$, which approximately meets the target of de-mixing. Note that interpolation-based
Algorithm 1 PAD Using DF-DM

Require: Feature Extractor $D(\cdot)$; Generator $G(\cdot)$; Discriminator $F(\cdot)$; Training Set $X_i$; PA Detector $H(\cdot)$;
Ensure: $1$: Trained $H(\cdot)$;
$2$: while $D(\cdot)$ has not converged do
$3$: for $x_i, x_j$ in $X_i$ do
$4$: Derive folded input $f_i$ from $T(x_i)$;
$5$: Reconstruct $f_i$ to $y_i$ through $G(D(f_i))$;
$6$: Update $G(\cdot)$ and $D(\cdot)$ by minimizing $I(1)$;
$7$: Update $F(\cdot) by maximizing $L^g$;
$8$: Calculate $L_{ij}$ from $x_i$ and $x_j$ through (2);
$9$: Update $D(\cdot)$ by minimizing $L^d$;
end for
end while
$11$: while $H(\cdot)$ has not converged do
$12$: Adopt $D(\cdot)$ as the initialization of $H(\cdot)$;
$13$: for $x_i$ in $X_i$ do
$14$: Obtain spoofness score $v_i$ by $H(x_i)$;
$15$: Update $H(\cdot)$ by minimizing (5);
end for
end while
$18$: end
$19$: Return $H(\cdot)$;

consistency has a trivial solution, e.g., embedding the same code for all the images, $\epsilon \in \mathcal{N}(0, 1)$ is thus added into $L_i$ to enhance the gradients against collapsing cases.

C. DF-DM-Based PAD

Considering the complementarity between de-folding and de-mixing, the proposed method trains $D(\cdot)$ with both the pretext tasks simultaneously, and the total learning objective can be concluded as

$$
\min_{G, D, f_i \sim T(X)} \mathcal{L}_f(y_i, x_i) + \mathcal{L}_g(y_i, x_i) + \mathcal{L}_t(x_i, x_j).
$$

(4)

After training, $D(\cdot)$ is used as the initialization for a PA detector $H(\cdot)$. Compared with $D(\cdot)$, $H(\cdot)$ has an additional fully connected layer to map $z_i$ into a single scalar $v_i$, i.e., spoofness score. The spoofness score reflects the category probability (PA or not) of the given sample $x_i$. $H(\cdot)$ is trained through a common cross-entropy-based objective as follows:

$$
\mathcal{L}_c(x_i, v_i) = - \mathbb{E}_{x_i \in X_i} \log(v_i) + (1 - u_i) \log(1 - v_i)
$$

(5)

where $v_i = H(x_i)$, and $u_i$ is the category annotation of $x_i$. For clarity, the proposed method is summarized in Algorithm 1.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the performance of the proposed method, extensive experiments are carried on the publicly available datasets, including LivDet2017 [62], OULU-NPU [67], CASIA-FASD [68], Idiap Replay-Attack [69], MSU-MFSD [70], Rose-Youtu [71], [72], and WMCA [73]. We first introduce the datasets and the corresponding implementation details.

Then, the effectiveness of the proposed method is validated by analyzing the contribution of each component. Since this is the first time to adopt self-supervised learning for PAD, we finally compare the proposed method with both the existing self-supervised methods and PA detectors to further prove the superiority of the proposed method.

A. Datasets and Implementation Details

As the proposed method is evaluated in two modalities, including fingerprint and face, we separately introduce the details of the corresponding protocols as follows.

1) Fingerprint: Due to the complete experimental settings, LivDet 2017 [62] is used to test the methods on fingerprint PAD. Table A.1 in the appendix summarizes the information of LivDet2017, which is used to evaluate the performance of fingerprint PAD. The dataset comprises over 17,500 fingerprint images captured from three readers, i.e., Green Bit, Orcanthus, and Digital Persona. Specifically, GreenBit is used for Italian border controls and insurance of Italian electronic documents. Orcanthus is widely used in the personal computer (PC). And Digital Persona is adopted in mobile devices, such as Nexus 7 tablet. Hence, the adopted readers are reliable for testing the practical performance of fingerprint PAD. For each sensor, about 1760 fingerprint images are used for training, 440 images for validation, and 3740 images for testing.

To evaluate the generalization of the competing methods, cross-material and cross-sensor settings [4] are used in this article. For cross-material cases, the spoof materials available in the test set are deemed as unknown materials, which are inaccessible during training. The partition of materials follows the setting in [4]. In the cross-sensor protocol, PA detectors are trained by the images collected using a randomly selected sensor and then tested using the images from the other sensors. EER, area under curve (AUC), and TDR @ FDR = 1% are used to evaluate the performance of detection.

In terms of the network architecture, MobileNet V2 [74] is selected as the backbone for the feature extractor and discriminator, while the corresponding generator is designed by following the U-Net architecture [75]. Note that to test the capacity of feature extraction and reduce the dependence on data scale, only the training set adopted for PAD is drawn to train the proposed method.

We compare the proposed method with both the self-supervised learning-based methods and PA detectors. For the self-supervised learning-based methods, GAN-based discriminator [76] and auto-encoder-based encoder [76] are set as the baseline of generative learning, while MoCo V2 [55] is selected as the representative method of contrastive learning. Regarding PA detector, LivDet 2017 winner [62] and FSB [21] are adopted as the competing method. For a more comprehensive analysis of the proposed method, multiple model-based PA detectors, including RTK-PAD [4] and FSB + UMG Wrapper [9], are also included for reference.

2) Face: To test the performance of face PAD, four datasets (details have shown in the appendix), including OULU-NPU [67] (denoted as O), CASIA-FASD [68] (denoted as C), Idiap Replay-Attack [69] (denoted as I), and MSU-MFSD [70] (denoted as M), are adopted in this article for evaluation using
two cross-dataset protocols. In Protocol-1, [O, M] and [C, I] are set as two groups, and the model is trained in one group and tested in the other. While in Protocol-2, three datasets are used to train, and the other dataset is adopted for evaluation. In these two protocols, we use the whole dataset in O, C, I, and M to train and test. For each video, only one randomly selected frame is used to train or test the detectors. In particular, printed photographs, display photographs, and replayed videos are used to attack the facial recognition systems. Various acquisition devices, such as laptops and smartphones, are considered in different datasets. Hence, the robustness and generalization of the PAD methods can be tested through the cross-dataset protocol.

In this case, the MTCNN algorithm [78] is adopted for face detection and alignment. All the detected faces are resized to (256,256). ResNet18 [79] is set as the backbone for the feature extractor and discriminator.

Besides the mentioned self-supervised methods, the state-of-the-art PA detectors, including DeepPixBiS [77], SSDG-R [18], and CDC [38], are conducted in this article, the results are shown in Table II. The effectiveness is validated by the improvement of such methods adopting the proposed method as initialization. We also compare our method with other self-supervised learning based method, including GAN based Discriminator [76], AE based Encoder [76], MoCo V2 [55]. The experimental results are shown in Table III. And MS-LBP [80], Binary CNN [81], IDA [70], Color Texture [82], LBP-TOP [83], Auxiliary [7], and MaddG [84] are set as the baselines for reference.

This article adopts the public platform PyTorch for all the experiments using a workstation with CPUs of 2.8 GHz, RAM of 512 GB, and GPUs of NVIDIA Tesla V100.

### B. Effectiveness Analysis of the Proposed Method

To quantify the contribution of de-mixing and de-folding, we test the performance of PAD with or without the corresponding pretext task. Tables IV and V show the results on fingerprint and face cases, respectively. The corresponding ROC curves are shown in the appendix. In face PAD, we follow [41] and use only protocol-1 for the ablation study. There are many more unseen PAs than the known ones in the training set. Using three datasets for training and one dataset for testing is not strict to the real application scenario. Hence, we use protocol-1 for our experiments. The numbers of data in O, C, I, and M are entirely different. To make the number of the training set and test set as close as possible, O and M are assigned to the same group. C and I are assigned to another group in protocol-1. Compared with protocol-2, protocol-1 uses fewer data for training, which makes the convergence of the network more difficult. It can show the generalization ability and effectiveness of our method.

The baseline is set as the model pretrained from ImageNet for PAD. Compared with the baseline, both de-folding and de-mixing can provide more reasonable initialization. Specifically, an increase of 9.58% in mean TDR@FDR = 1.0% is achieved by adopting de-Mixing as the pretext task in Table IV. When it comes to face, de-folding improves the EER of baseline from 26.90% to 22.86%. This indicates that both the components in the proposed method can promote PAD effectively. Among all the cases, the most significant improvement is obtained when all the designed components are adopted, i.e., DF-DM can reach 18.78% and 2.59% mean EER in the face and fingerprint cases, respectively. The corresponding ROC curves are shown in the appendix. In face PAD, we follow [41] and use only protocol-1 for the ablation study. There are many more unseen PAs than the known ones in the training set. Using three datasets for training and one dataset for testing is not strict to the real application scenario. Hence, we use protocol-1 for our experiments. The numbers of data in O, C, I, and M are entirely different. To make the number of the training set and test set as close as possible, O and M are assigned to the same group. C and I are assigned to another group in protocol-1. Compared with protocol-2, protocol-1 uses fewer data for training, which makes the convergence of the network more difficult. It can show the generalization ability and effectiveness of our method.

The baseline is set as the model pretrained from ImageNet for PAD. Compared with the baseline, both de-folding and de-mixing can provide more reasonable initialization. Specifically, an increase of 9.58% in mean TDR@FDR = 1.0% is achieved by adopting de-Mixing as the pretext task in Table IV. When it comes to face, de-folding improves the EER of baseline from 26.90% to 22.86%. This indicates that both the components in the proposed method can promote PAD effectively. Among all the cases, the most significant improvement is obtained when all the designed components are adopted, i.e., DF-DM can reach 18.78% and 2.59% mean EER in the face and fingerprint cases, respectively. The corresponding ROC curves are shown in the appendix. In face PAD, we follow [41] and use only protocol-1 for the ablation study. There are many more unseen PAs than the known ones in the training set. Using three datasets for training and one dataset for testing is not strict to the real application scenario. Hence, we use protocol-1 for our experiments. The numbers of data in O, C, I, and M are entirely different. To make the number of the training set and test set as close as possible, O and M are assigned to the same group. C and I are assigned to another group in protocol-1. Compared with protocol-2, protocol-1 uses fewer data for training, which makes the convergence of the network more difficult. It can show the generalization ability and effectiveness of our method. The baseline is set as the model pretrained from ImageNet for PAD. Compared with the baseline, both de-folding and de-mixing can provide more reasonable initialization. Specifically, an increase of 9.58% in mean TDR@FDR = 1.0% is achieved by adopting de-Mixing as the pretext task in Table IV. When it comes to face, de-folding improves the EER of baseline from 26.90% to 22.86%. This indicates that both the components in the proposed method can promote PAD effectively. Among all the cases, the most significant improvement is obtained when all the designed components are adopted, i.e., DF-DM can reach 18.78% and 2.59% mean EER in the face and fingerprint cases, respectively. The corresponding ROC curves are shown in the appendix.

### V. Ablation Study

#### A. Comparison With Related Methods

1) Comparison With Self-Supervised Methods: Due to the difference between the natural and face/fingerprint images, directly adopting existing self-supervised methods for PAD is not a proper choice. Hence, we proposed a self-supervised learning-based method: DF-DM. The pipeline of self-supervised learning contains two training steps: self-supervised pretext task training and supervised downstream task training. In this article, DF-DM is self-supervised pretext tasks, and the PAD classification is the downstream task. In the pretext tasks, we do not use any PAD labels, and...
the pseudolabels we use are automatically generated from the images themselves. We only use PAD labels in downstream tasks. To validate the effectiveness of the proposed method, we compare DF-DM with the existing self-supervised methods. As the results listed in Table III, the proposed method outperforms the existing methods significantly. “Ours: DF-DM” and “Ours: DF-DM(ImageNet)” are training in self-supervised learning manner. “Ours: DF-DM” denotes that in the pretext task training step (DF-DM), the network parameters are trained from scratch. But for “Ours: DF-DM (ImageNet),” the network parameters are pretrained from ImageNet. “PreTrained from ImageNet” is training in a supervised learning manner, so it does not contain the DF-DM training step. And in the supervised learning step (PAD classification), the network parameters are pretrained from ImageNet. In terms of face, when trained from scratch, our method can achieve an EER of 31.28%, exceeding other self-supervised methods by around 4%–10% absolutely. Meanwhile, the proposed method can further improve the performance of the model pretrained from ImageNet. Typically, in the case of fingerprint, DF-DM reaches 90.96% TDR when FDR = 1.0%, which outperforms the initialization from ImageNet by a large margin, i.e., 90.96% versus 73.92%. Note that the data scale of PAD is limited; hence, directly using MoCo cannot reach competitive results and may lead the model to learn useful features for identification, but it is useless for PAD.

2) Comparison With PA Detectors: To further verify the effectiveness of the proposed method, we compare it with the state-of-the-art methods. As the results listed in Table VI, under the cross-material and cross-sensor settings, the proposed method can outperform other single-model-based methods by a large margin. In the cross-sensor case, compared with FSB, a reduction of 12.58% in average classification error (ACE) can be obtained by DF-DM. By comprehensively analyzing cross-material and -sensor protocols, DF-DM can promote PA detector to 11.15% mean ACE, even exceeding the multiple-model-based methods, which convincingly proves the advantage of DF-DM. Regarding the cross-material case, a 2.48% of ACE can be derived by our proposed method, which outperforms FSB + UMG Wrapper (4.12%) and is close to RTK-PAD (2.28%). More details of the cross-material and cross-sensor are shown in the appendix.

When it comes to face, we first conduct our proposed method in a famous publicly available benchmark to justify its effectiveness. As listed in Table VII, DF-DM can reach to 13.30% half total error rate (HTER) and 93.66% AUC. Without any changes in learning objectives and network architectures, a baseline model, ResNet-18, can directly surpass most competing methods by adopting DF-DM as the initialization. When combined with SSDG-R, the proposed method can achieve the best PAD performance, which improves SSDG-R from 11.29% HTER to 9.49%. EPCR [61] and PatchNet [59] in Table VII are the self-supervised based methods. Our proposed DF-DM can suppress EPCR [61] around 2.53% in HTER. Meanwhile, we reimplement some famous PA detectors and investigate the improvement from DF-DM in different detectors. As listed in Table II, DF-DM can facilitate detection performance by around 5%–25% in mean TDR@FDR = 1.0%. When DeepPixBiS is used as the detector, DF-DM can improve the AUC of PAD from 82.42% to 89.48%. The experimental results indicate that the proposed method is general and can be integrated with various PA detectors.

B. Intradataset Testing

We also conduct experiments on Rose-YouTu [71], [72] and WMCA [73] for intradataset testing. The WMCA dataset [73] contains a wide variety of 2-D and 3-D PAs, with a total of 1679 video samples from 72 subjects. The Rose-YouTu dataset [71], [72] covers a large variety of illumination conditions, camera models, and attack types, which consists of 4225 videos with 25 subjects in total. Attack presentation classification error rate (APCER), bona fide presentation classification error rate (BPCER), and ACE rate (ACER) are used for intradataset testing.

1) Results on WMCA: The WMCA dataset consists of 1941 short video recordings of both bona fide and PAs from

| Baseline | In-Image De-Folding | Out-of-Image De-Mixing | GreenBi | DigitalPersona | Or航运 | Mean ± s.d. |
|----------|---------------------|------------------------|---------|----------------|---------|------------|
|          |                     |                        | EER(%) | AUC(%) | TDR(%) | EER(%) | AUC(%) | TDR(%) | EER(%) | AUC(%) | TDR(%) | EER(%) | AUC(%) | TDR(%) |
| ✓        | ×                   | ×                      | 20.33  | 84.51  | 9.79   | 25.38  | 81.16  | 26.70  | 22.86  | 5.37   | 82.84  | 2.37   | 18.25  | 11.96  |
| ✓        | ×                   | ✓                      | 21.07  | 85.17  | 11.93  | 26.33  | 80.25  | 27.49  | 23.70  | 5.72   | 82.71  | 3.48   | 19.71  | 11.00  |
| ✓        | ✓                   | ✓                      | 18.96  | 89.48  | 30.48  | 18.60  | 89.76  | 30.30  | 18.78  | 0.25   | 89.62  | 0.20   | 30.39  | 0.13   |

TABLE IV
PERFORMANCE OF THE PROPOSED METHOD WITH OR WITHOUT EACH COMPONENT IN TERMS OF EER (%) ↓, AUC (%) ↑, AND TDR(%)@FDR = 1.0% ↑ UNDER THE CROSS-MATERIAL SETTING ON LivDet2017

| Baseline | In-Image De-Folding | Out-of-Image De-Mixing | [O,M] to [C,I] | [C,H] to [O,M] | Mean ± s.d. |
|----------|---------------------|------------------------|----------------|----------------|------------|
|          |                     |                        | EER(%) | AUC(%) | TDR(%) | EER(%) | AUC(%) | TDR(%) | EER(%) | AUC(%) | TDR(%) |
| ✓        | ×                   | ×                      | 26.90  | 79.10  | 0.06   | 11.37  | 10.32 |
| ✓        | ×                   | ✓                      | 22.86  | 82.84  | 2.37   | 18.25  | 11.96 |
| ✓        | ✓                   | ✓                      | 23.70  | 82.71  | 3.48   | 19.71  | 11.00 |

TABLE V
PERFORMANCE OF THE PROPOSED METHOD WITH OR WITHOUT EACH COMPONENT IN TERMS OF EER (%) ↓, AUC (%) ↑, AND TDR(%)@FDR = 1.0% ↑ UNDER THE CROSS-DATASET SETTING ON OULU-NPU (O), CASIA-FASD (C), IDIAP REPLAY-ATTACK (I), AND MSU-MFSD (M) (Protocol-1)
72 identities. The data are recorded from several channels, including color, depth, infrared, and thermal. We randomly select one frame from each video for training and testing. Hence, the training and testing sets contain a total of 1941 images. The experiment results on Rose-Youtu are listed in Table VIII. DF-DM can achieve 2.84% ACER, exceeding baseline performance by 3.76%. The combination of SSDG-R with our method gets a better result, which improves the baseline from 6.60% ACER to 0.89% ACER.

2) Results on Rose-Youtu: The Rose-Youtu dataset consists of 4225 videos with 25 subjects in total. For each video, we use the same method as WMCA to extract images, and the dataset contains a total of 4255 images. The experiment results on Rose-Youtu are listed in Table IX. Our proposed DF-DM outperforms the baseline method by around 0.53% in ACER. When combined with SSDG-R, the proposed method gets the best results.

C. Training With Different Strategies

The proposed method can be trained with different strategies, including step and together training. In the step training, it first uses the self-supervised learning task DF-DM to train the model without using PAD labels and then fine-tune the network parameters in a supervised training manner with PAD labels. In the together training, the two training steps, the DF-DM training without PAD labels and supervised training with PAD labels, are trained simultaneously. The experimental results are shown in Table X, which indicates that step training outperforms together training by a wide margin.

VI. VISUALIZATION RESULTS

To further investigate the advance of the proposed method over other baselines, we visualize the discriminative features extracted by the models with the same architecture but different initialization. Besides, to further clarify the contribution of DF-DM for generalization, we then visualize the discriminative features among different datasets. The visualization results are presented in the appendix.
VII. CONCLUSION

In this article, we proposed a self-supervised learning-based method to improve the generalization performance of PA detectors. The de-folding and de-mixing pretext tasks included in the method work together as a local–global strategy. That is, de-folding requires the model to reconstruct the folded image to the row by extracting region-specific features, i.e., local information, while de-mixing drives the model to derive instance-specific features, i.e., global information, by disentangling the mixed samples. The generalization ability is finally improved by the comprehensive local–global view of training samples. The effectiveness of the proposed method is verified in terms of face and fingerprint PAD, including seven publicly available datasets: LivDet2017, OULU-NPU, CASIA-FASD, Idiap Replay-Attack, MSU-MFSD, WMCA, and Rose-YouTu.

In the future, we will further investigate the application of the proposed method in other tasks, such as fingerprint/face recognition and face detection/alignment.

REFERENCES

[1] R. Raghavendra and C. Busch, “Presentation attack detection methods for face recognition systems: A comprehensive survey,” Comput. Surv., vol. 50, no. 1, pp. 1–37, Mar. 2017.

[2] J. M. Singh, A. Madhun, G. Li, and R. Ramachandra, “A survey on unknown presentation attack detection for fingerprint,” 2020, arXiv:2005.08337.

[3] 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, Jul. 2020.

[4] H. Liu, W. Zhang, F. Liu, H. Wu, and L. Shen, “Fingerprint presentation attack detector using global-local model,” IEEE Trans. Cybern., vol. 52, no. 11, pp. 1–14, Jun. 2021.

[5] F. Liu, H. Liu, W. Zhang, G. Li, and L. Shen, “One-class fingerprint presentation attack detection using auto-encoder network,” IEEE Trans. Image Process., vol. 30, pp. 2394–2407, 2021.

[6] R. Raghavendra, K. B. Raja, and C. Busch, “Presentation attack detection for face recognition using light field camera,” IEEE Trans. Image Process., vol. 24, no. 3, pp. 1060–1075, Mar. 2015.

[7] 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.

[8] K. Larbi, W. Ouarda, H. Drira, B. Ben Amor, and C. Ben Amar, “DeepColorFASD: Face anti-spoofing solution using a multi channelled color spaces CNN,” in Proc. BMC, Oct. 2018, pp. 4011–4016.

[9] T. Chugh and A. K. Jain, “Fingerprint spoof detector generalization,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 42–55, 2021.

[10] Y. Wang, X. Song, T. Xu, Z. Peng, and X.-J. Wu, “From RGB to depth: Domain transfer network for face anti-spoofing,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4280–4290, 2021.

[11] J. Guo, X. Zhu, J. Xiao, Z. Lei, G. Wan, and S. Z. Li, “Improving face anti-spoofing by 3D virtual simulation,” in Proc. ICB, Jun. 2019, pp. 1–8.

[12] W. Zhang et al., “A multi-task network with weight decay skip connection training for anomaly detection in retinal fundus images,” in Medical Image Computing and Computer Assisted Intervention. Singapore: Springer, Sep. 2022, pp. 656–666.

[13] T. de Freitas Pereira, A. Anjou, J. M. De Martino, and S. Marcel, “Can face anti-spoofing countermeasures work in a real world scenario?” in Proc. ICB, Jun. 2013, pp. 1–8.

[14] J. Yang, Z. Lei, S. Liao, and S. Z. Li, “Face liveness detection with component dependent descriptor,” in Proc. ICB, Jun. 2013, pp. 1–6.

[15] P. Patel, H. Han, and A. Jain, “Secure face unlock: Spoof detection on smartphones,” IEEE Trans. Inf. Forensics Security, vol. 11, no. 10, pp. 2268–2283, Jun. 2016.

[16] Z. Boulkenafet, J. Komulainen, and A. Hadid, “Face anti-spoofing using speeded-up robust features and Fisher vector encoding,” IEEE Signal Process. Lett., vol. 24, no. 2, pp. 141–145, Feb. 2016.

[17] Z. Wang, B. Du, and Y. Guo, “Domain adaptation with neural embedding matching,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 7, pp. 2387–2397, Jul. 2019.

[18] Y. Jia, J. Zhang, S. Shan, and X. Chen, “Single-side domain generalization for face anti-spoofing,” in Proc. CVPR, Jun. 2020, pp. 8484–8493.

[19] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.

[20] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, “Fingerprint liveness detection using convolutional neural networks,” IEEE Trans. Inf. Forensics Security, vol. 11, no. 6, pp. 1206–1213, Jun. 2016.

[21] T. Chugh, K. Cao, and K. Anil Jain, “Fingerprint spoof buster: Use of minutiae-centered patches,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 9, pp. 2190–2202, Sep. 2018.

[22] J. A. Pereira, A. F. Sequeira, D. Pernes, and J. S. Cardoso, “A robust fingerprint presentation attack detection method against unseen attacks through adversarial learning,” in Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), Sep. 2020, pp. 1–5.

[23] F. Liu, Z. Kong, H. Liu, W. Zhang, and L. Shen, “Fingerprint presentation attack detection by channel-wise feature denoising,” IEEE Trans. Inf. Forensics Security, vol. 17, pp. 2963–2976, 2022.

[24] A. F. Sequeira and J. S. Cardoso, “Fingerprint liveness detection in the presence of capable intruders,” Sensors, vol. 15, no. 6, pp. 14615–14638, Jun. 2015.

[25] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, “Face anti-spoofing with human material perception,” in Proc. Eur. Conf. Comput. Vis. Glasgow, U.K.: Springer, 2020, pp. 557–575.

[26] Y. Qin et al., “Learning meta model for zero-and few-shot face anti-spoofing,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 11916–11923.

[27] Z. Wang et al., “Deep spatial gradient and temporal depth learning for face anti-spoofing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020, pp. 5042–5051.

[28] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, “NAS-FAS: Static-dynamic central difference network search for face anti-spoofing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 9, pp. 3005–3023, Sep. 2020.

[29] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, “Revisiting pixel-wise supervision for face anti-spoofing,” IEEE Trans. Biometrics, Behav., Identity Sci., vol. 3, no. 3, pp. 285–295, Jul. 2021.

[30] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao, and Z. Lei, “Meta-teacher for face anti-spoofing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 52, no. 11, pp. 12315–12328, Nov. 2020.

[31] Z. Yu, X. Li, P. Wang, and G. Zhao, “TransRPG: Remote photoplethysmography transformer for 3D mask face presentation attack detection,” IEEE Signal Process. Lett., vol. 28, pp. 1290–1294, 2021.

[32] Z. Wang et al., “Domain generalization via shuffled style assembly for face anti-spoofing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2022, pp. 4123–4133.

[33] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, “Dual-cross central difference network for face anti-spoofing,” 2021, arXiv:2105.01290.

[34] Z. Yu, R. Cai, Z. Li, W. Yang, J. Shi, and A. C. Kot, “Benchmarking joint face spoofing and forgery detection with visual and physiological cues,” 2022, arXiv:2208.05401.

[35] Y. Cui, Z. Yu, W. Peng, and L. Liu, “Rethinking few-shot class-incremental learning with open-set hypothesis in hyperbolic geometry,” 2022, arXiv:2207.09963.

[36] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, “Deep tree learning for zero-shot face anti-spoofing,” in Proc. CVPR, Jun. 2019, pp. 4680–4689.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Wentian Zhang received the B.S. degree from the Shaanxi University of Science and Technology, Xi’an, China, in 2020. He is currently pursuing the M.S. degree with Shenzhen University, Shenzhen, China.

His research interests include image processing, anomaly detection, and graph embedding.

Feng Liu (Member, IEEE) received the B.Sc. and M.Sc. degrees from Xidian University, Xi’an, Shaanxi, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2014.

She is currently an Associate Professor with the School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. She has published more than 40 papers in academic journals and conferences and participated in many research projects either as a principal investigator or as a primary researcher. Her research interests include pattern recognition and image processing, especially for applications to fingerprints.

Dr. Liu is a reviewer of many renowned field journals and conferences.

Wenhan Luo (Member, IEEE) received the B.E. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.E. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012, and the Ph.D. degree from the Imperial College London, London, U.K., in 2016.

He is currently an Associate Professor with Sun Yat-sen University, Guangzhou, China. Prior to that, he has worked as the Research Scientist of Tencent, Shenzhen, Guangdong, China, and Amazon, California, USA. He has published more than 40 papers in top conferences and leading journals, including International Conference on Machine Learning (ICML), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), Annual Meeting of the Association for Computational Linguistics (ACL), AAAI Conference on Artificial Intelligence (AAAI), International Conference on Learning Representations (ICLR), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), International Journal of Computer Vision (IJCV), Artificial Intelligence (AI), and IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP). His research interests include several topics in computer vision and machine learning, such as image/video synthesis, image/video quality restoration, and reinforcement learning.

Dr. Luo has been a reviewer, a senior PC member, and a guest editor of several prestigious journals and conferences.

Haozhe Liu received the M.S. degree in computer science from Shenzhen University, Shenzhen, China, in 2022. He is a graduate student from Shenzhen University in 2022.

He has coauthored several papers on top-tier journals and conferences, including IEEE Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), Medical Image Computing and Computer Assisted Intervention (MICCAI), and IEEE TRANSACTIONS ON IMAGE PROCESSING. His research interests include regularization, self-supervised learning, adversarial learning, and reinforcement learning.

Mr. Liu serves as a Reviewer of the top-tier conferences, e.g., CVPR 2022, ICML 2022, ECCV 2022, and MICCAI 2022.

Linlin Shen (Senior Member, IEEE) received the B.Sc. and M.Eng. degrees from Shanghai Jiaotong University, Shanghai, China, in 1997 and 2000, respectively, and the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 2005.

He was a Research Fellow with the University of Nottingham, working on MRI brain image processing. He is currently a Pengecheng Scholar Distinguished Professor with the School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is also an Honorary Professor with the School of Computer Science, University of Nottingham, and a Distinguished Visiting Scholar with the University of Macao, Macau, China. He serves as the Director of the Computer Vision Institute, AI Research Center for Medical Image Analysis and Diagnosis and China–U.K., a joint Research Laboratory for visual information processing. His research interests include deep learning, facial recognition, analysis/synthesis, and medical image processing.

Dr. Shen is listed as the Most Cited Chinese Researchers by Elsevier. He received the Most Cited Paper Award from the Image and Vision Computing journal. His cell classification algorithms were the winners of the International Contest on Pattern Recognition Techniques for Indirect Immunofluorescence Images held by International Conference on Image Processing (ICIP) 2013 and International Conference on Pattern Recognition (ICPR) 2016.

Linlin Shen

Raghavendra Ramachandra (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from the University of Mysore, Mysuru, India, and the Institute Telecom, and Telecom Sudparis, Evry, France (carried out as collaborative work), in 2010.

He was a Researcher with the Istituto Italiano di Tecnologia, Genoa, Italy, where he has worked with video surveillance and social signal processing. He is currently a Full Professor with the Institute of Information Security and Communication Technology (IIK), Norwegian University of Science and Technology (NTNU), Gjøvik, Norway. He has authored several papers and is a reviewer of several international conferences and journals. He also holds several patents in biometric presentation attack detection and morphing attack detection. His main research interests include deep learning, machine learning, data fusion schemes, and image/video processing, with applications to biometrics, multimodal biometric fusion, human behavior analysis, and crowd behavior analysis.

Dr. Ramachandra has also been involved in various conference organizing and program committees and has served as an associate editor for various journals. He has received several best paper awards. He has participated (as a PI, a co-PI, or a contributor) in several EU projects, IARPA USA, and other national projects. He has been serving as an Editor for the ISO/IEC 24722 standards on multimodal biometrics and an active contributor to the ISO/IEC SC 37 standards on biometrics.