Dense Hybrid Attention Network for Palmprint Image Super-Resolution

Yao Wang, Lunke Fei, *Senior Member, IEEE*, Shuping Zhao, *Member, IEEE*, Qi Zhu, Jie Wen, *Senior Member, IEEE*, Wei Jia, *Member, IEEE*, and Imad Rida

***Abstract*—Palmprint has attracted increasing attention for biometric recognition in recent years due to its outstanding reliability, user-friendliness and hygiene. However, existing palmprint recognition methods usually require high-quality palmprint images with clear texture and line patterns; however, in practical applications palmprint images are usually of low quality. In this study, we propose a dense hybrid attention (DHA) network for palmprint image super-resolution (SR) by recovering the clear palmprint-specific characteristics. The proposed DHA network first obtains the high-dimensional shallow representation via a single convolution layer, and then jointly learns the local and global palmprint-specific features via parallel convolutional neural network (CNN)- and transformer-based branches. Particularly, we develop two enhanced spatial and channel attention (CA) modules to adaptively emphasize the local position-specific characteristics of palmprints, such that the SR palmprint images can be well recovered with clear texture and edge characteristics. Experimental results on three publicly used palmprint databases clearly show the effectiveness of the proposed method for palmprint image SR.**

***Index Terms*—Attention mechanism, multibranch hybrid model, palmprint images, super-resolution (SR).**

I. Introduction

BIOMETRIC recognition has been widely used for human authentication [1], and various biometric traits, such as fingerprint, face, and gait, have been successfully developed for practical biometric recognition systems. In recent years, palmprint has received considerable research interest for biometric recognition due to its several advantages: rich unique and discriminative features, noninvasiveness, low-resolution (LR) imaging, user-friendliness, and contactless collection with high hygiene. In the literature, various methods have been proposed for palmprint recognition, such as local feature encoding [1], subspace representation [2], and deep learning [3] methods. It is noted that most existing palmprint descriptors usually first crop the regions of interest (ROI) of the palmprint images and afterward learn and extract the discriminative features from the ROIs, such that the performance of palmprint recognition heavily relies on the quality of the original acquired palmprint samples. For this reason, to maintain palmprint image quality, they are usually captured by a uniform camera under a constrained background for reliable palmprint recognition [4], [5], [6]. However, in practical application scenarios, palmprint images are possibly captured by using different kinds of cameras with different resolutions under varying background environments with various illumination changes. Therefore, the palmprint images in practical applications are easily affected by the imaging hardware, contactless acquisition manner and complex environment. Therefore, improving the quality of palmprint samples plays a critical role and remains a challenging problem for reliable palmprint recognition.

Image super-resolution (SR) has been demonstrated its promising effectiveness for image perceptual quality improvement by recovering the high-resolution (HR) images from their LR, which plays a significant role in various fields, such as biometric imaging, surveillance, and entertainment for accurate analysis and detection. In the literature, there has been a variety of SR methods, which can be roughly categorized into interpolation-based, reconstruction-based and learning-based methods. In general, the interpolation-based methods produce an HR sample from an LR image by calculating the pixel intensities on an upsampled grid, and the typical methods include local variance coefficients [7] and adaptive structure kernels [8]. For example, Zhang et al. [9] introduced a bivariate rational fractal interpolation SR method by learning the analytical features based on various scaling factors and shapes. In contrast, the reconstruction approaches assume that LR images are essentially developed from the degradations, such as warping and blurring, and thus they usually require different priors, such as edge directed priors [10] and similarity redundancy [11], to recover the SR image by fitting degradation functions. Moreover, due to the outstanding performance

Manuscript received 15 October 2023; accepted 15 December 2023. Date of publication 10 January 2024; date of current version 19 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176066, Grant 62106052, and Grant 62006059; and in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515012717. This article was recommended by Associate Editor B. Zhang. *(Corresponding author: Lunke Fei.)*

Yao Wang, Lunke Fei, and Shuping Zhao are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: wangyaofzx@163.com; flksxm@126.com; yb77458@um.edu.mo).

Qi Zhu is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: zhuqi@nuaa.edu.cn).

Jie Wen is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: jiewen_pr@126.com).

Wei Jia is with the School of Computer and Information, Hefei University of Technology, Hefei 230002, China (e-mail: china.jiawei@139.com).

Imad Rida is with the Centre de Recherches de Royallieu, Université de Technologie de Compiègne, 60200 Compiègne, France (e-mail: imad.rida@utc.fr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2023.3344607.

Digital Object Identifier 10.1109/TSMC.2023.3344607

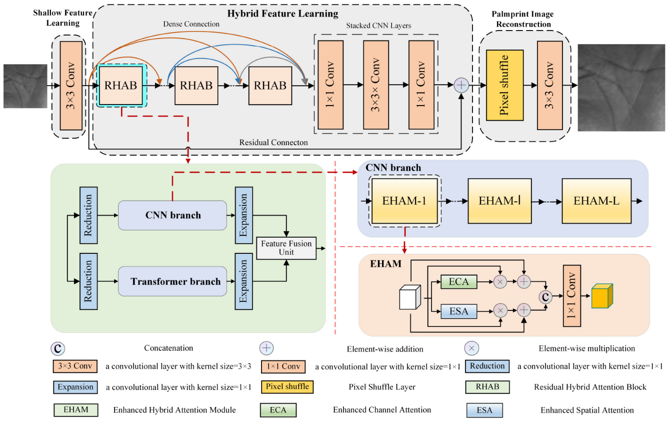


Fig. 1. Main framework of the proposed DHA network is composed of shallow feature learning, hybrid feature learning, and palmprint image reconstruction subnetworks.

of deep learning in image analysis, recently, there have been an increasing number of deep learning-based SR methods in [12], [13], and [14]. For example, Dong et al. [12] proposed a end-to-end mapping convolutional network to convert a LR image into its HR representation. Zhang et al. [13] proposed a residual channel attention network (RCAN) by embedding the channel attention (CA) mechanism into a convolutional framework for image SR. Chen et al. [14] built a pure ViT-based network to track the image SR tasks. While most existing SR methods have achieved promising performance for common image SR, they cannot be directly applied to palmprint image SR due to the intrinsic characteristics of palmprints, such as palmar ridge patterns and palmar flexion creases. To address this, this article specifically concentrates on studying palmprint image SR.

In this article, we propose a dense hybrid attention (DHA) network for palmprint image SR. Fig. 1 shows the basic framework of the proposed DHA network, which consists of shallow feature learning, hybrid feature learning and palmprint image reconstruction subnetworks. First, the shallow feature learning subnetwork converts the palmprint images into a high-dimensional shallow representation via a single 3×3 convolution layer. Then, the hybrid feature learning subnetwork uses multiple residual hybrid attention blocks (RHABs) with dense and residual connections to learn hierarchical palmprint image features, where parallel convolutional neural network (CNN) and transformer learning branches are embedded into the RHAB to simultaneously learn the local and global characteristics of palmprint images. Particularly, we develop an enhanced channel attention (ECA) and an enhanced spatial attention (ESA) into the CNN branch to specifically learn the intrinsic texture and edge information of palmprints. Last, based on both shallow and high-level palmprint feature maps, the image reconstruction subnetwork recovers the SR palmprint image via a PixelShuffle layer [15] and a convolution layer.

The main contributions of this article are summarized as follows.

1) We introduce a DHA network for palmprint image SR by first converting a palmprint image into a high-dimensional shallow representation and then jointly learning the hierarchical local and global features of palmprint images. To the best of our knowledge, this work is the first study to focus on palmprint image SR.
2) Unlike existing standard channel and spatial attentions, we develop an ECA and an ESA to specifically learn the position-specific and pattern-specific palmprint features such that the palmprint images can be recovered into SR representations with clearer palmprint texture and edge characteristics.
3) We conduct extensive experiments on numerous palmprint benchmark datasets, and the comparative experimental results clearly show the superiority of the proposed DHA network.

The remainder of this article is organized as follows. Section II briefly reviews the related topics. Section III details our proposed DHA network. Section IV analyzes the experimental results. Section V concludes this article.

II. Related Work

A. Palmprint Images

The palm surface of humans consists of palmar friction ridges and flexion creases. In general, the friction ridges of palmprints are of small sizes, which usually require HR cameras to capture the ridge characteristics of the palmprint. As a result, the ridge-based characters, including the ridge patterns, ridge valleys, and ridge-ending minutia points, can only be seen in HR palmprint images with more than 500 PPI. Due to the expensive cost of acquisition and processing, HR palmprint images are usually used in legal and forensic applications. It is noted that a palmprint has a large skin area when compared with a fingerprint. Therefore, palmprint images can also be collected by using an inexpensive camera with low resolution (approximately 100 PPI), and existing palmprint recognition studies focus mostly on LR palmprint images for commercial or civil applications. Unlike a HR palmprint image that can clearly describe the ridge information, a LR palmprint image depicts the texture patterns formed by the palmar ridges, adding the line characteristics formed by the palmar creases. To extract the discriminant features from LR palmprint samples, a number of palmprint recognition methods have been successfully proposed in recent years [2], [16], [17], [18], while the existing methods usually concentrate only on feature extraction, learning and representation of LR palmprint samples. The final performance of palmprint recognition depends most on the LR palmprint image quality. In the literature on palmprint recognition, palmprint images are usually captured in constrained conditions, such as by using a uniform camera and under a uniform background, such that the palmprint image quality can be guaranteed. However, in real-world scenarios, palmprint images are possibly captured by using different types of cameras under complex backgrounds, making their quality low. To address this, this study focuses on palmprint image SR, aiming to improve the perceptual quality of palmprint images with clear palmprint-specific characteristics from LR palmprint samples.

B. Deep Learning for Image SR

Image SR aims to recover HR images from LR samples to upgrade the image perceptual quality for a variety of computer vision tasks, such as medical imaging and identity recognition [19]. Over the past decades, due to its tremendous achievement in computer vision tasks, a number of deep learning-based methods have been proposed for image SR and achieved promising performance. For instance, Dong et al. [12] designed a CNN-based image SR network and achieved superior performance over conventional learning methods. Inspired by this, numerous CNN-based networks [20], [21], [22], [23], [24], [25] were proposed to further increase the reconstruction performance for image SR. After that, the typical recursive neural network [21] and graph neural network [22] have also been applied for image SR. For instance, Kim et al. [23] designed a deep residual learning network by building more than 20 convolution layers to exploit abundant contextual information for image SR. Lim et al. [20] proposed an enhanced deep residual networks for single image super-resolution (EDSR) by stacking the modified residual blocks to expand the model size for better-SR performance, demonstrating the effectiveness of depth for image SR. Furthermore, Tai et al. [24] proposed a MemNet by embedding dense blocks into the convolution layers to explicitly mine persistent shallow features for building faithful SR images. In addition, with the rapid development of attention mechanisms, several recent studies [26], [27] have introduced attention mechanisms and transformers for channel- and spectral-specific feature learning to track SR tasks. For example, Kim et al. [26] developed a residual attention SR framework by applying the residual blocks and spatial CA to specifically learn the interchannel and intrachannel correlations, and Chen et al. [27] proposed a hybrid attention transformer (HAT) network to simultaneously exploit the global information and improve the representative ability for image SR. In this work, we develop a new hybrid attention network focusing on palmprint image SR by employing both a CNN and a transformer to capture multitype palmprint-specific features.

C. Attention Mechanism for Image SR

Inspired by the fact that human vision systems can effectively divert most attention to the important regions of images, attention mechanisms are introduced into deep learning networks to adaptively weight features based on sample importance. Attention mechanisms have been successfully used for a variety of visual tasks, such as object detection, semantic segmentation, and image SR. Motivated by the major success of attention mechanisms in special feature learning, in recent years, researchers have recently proposed attention-based learning for image SR, including CA [13], spatial attention [28], nonlocal attention [29], [30], self-attention [31], [32], and multiattention [33], [34] methods. For example, Zhang et al. [13] built a RCAN network for image SR by incorporating a CA mechanism into a simplified residual backbone. Liu et al. [28] designed an ESA network to specifically learn the local spatial information for image SR. To apply the uneven distribution of information

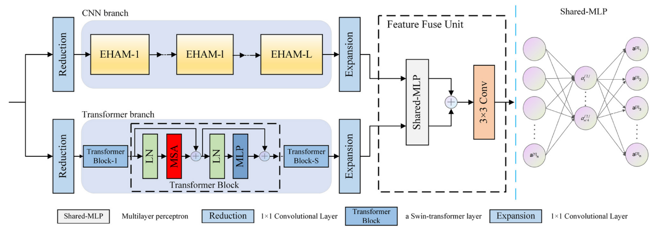


Fig. 2. Basic architecture of the RHAB, which consists of a CNN branch for local feature learning, a transformer branch for global feature learning, and an FFU for feature fusion of the double branches.

in corrupted images, a number of nonlocal attention-based methods, such as nonlocal recurrent network [30] and nonlocal sparse attention network (NLSA) [35], have been proposed to capture long-distance spatial information for image SR. Recently, by encoding distant dependencies and capturing global interactions, transformers based on self-attention have also been employed for image SR, and the typical methods include encoder–decoder-based transformer (EDT) [31], swin transformer-based image restoration (SwinIR) [36], aggregate enriched features extracted from both cnn and transformer (ACT) [27] and HAT [27]. For example, Liang et al. [36] presented a SwinIR framework based on the Swin Transformer [37] to build long-range dependency mechanism for image SR. Additionally, Niu et al. [34] presented a multiple attention mechanism-based SR network, i.e., holistic attention network (HAN) by employing CA and spatial attention to reconstruct more faithful SR images. In this study, we present a new hierarchical attention-based network for palmprint image SR by enhancing and combining spatial and CA to capture multiple palmprint features.

III. Proposed DHA Network

A. Architecture of the DHA Network

The proposed DHA consists of three subnetworks, as shown in Fig. 1, including shallow feature learning, hybrid feature learning, and palmprint image reconstruction subnetworks. Let $I_{\mathrm{LR}} \in \mathbb{R}^{C_{\mathrm{in}} \times H \times W}$ be the input LR palmprint image, where C_{in}, H and W denote the channel, height and width of the input sample, respectively. The shallow feature learning subnetwork employs a single 3×3 convolution layer to extract raw pixel-level shallow features, as follows:

$$F_0 = W_0 * I_{\mathrm{LR}} \tag{1}$$

where W_0 represents the weights of the 3×3 convolution layer, and "*" denotes the convolution operation. With the single-layer convolutional subnetwork, the input palmprint image can be easily converted from the RGB channel space into a higher-channel dimension for shallow palmprint feature representation, referred to as $F_0 \in \mathbb{R}^{C \times H \times W}$ (C denotes the number of channels), such that more hierarchical features can be exploited from the input image, leading to a better-visual representation [38]. To this end, we first convert the palmprint images into the high-dimensional shallow representation, and then learn the local and global characteristics of palmprint images from the high-dimensional shallow representation.

Having obtained the high-dimensional shallow feature representation of the palmprint images, we further develop a hybrid feature learning subnetwork to emphasize the high-level semantic features, such as texture and edge patterns of palmprint images. The hybrid feature learning subnetwork consists of multiple RHABs and three stacked convolution layers with dense and residual connections. The hybrid feature learning procedure can be presented as follows:

$$F_{DF} = H_{SC}(H_{DC}([F_1, \ldots, F_B]) + F_0) \tag{2}$$

where the F_{DF} represents the high-level palmprint feature descriptor learned from the hybrid feature learning subnetwork. $[F_1, \ldots, F_B]$ indicates the feature maps learned by the RHABs, where B indicates the total number of RHABs. In this article, we embed 9 RHABs into the hybrid feature learning subnetwork (i.e., $B = 9$). H_{DC} represents the dense connection to maximize the information flow between RHAB blocks. "+" fuses the low-level shallow features with the high-level hybrid features via a summation operation, which can stabilize the training. $H_{SC}(\cdot)$ represents the mapping of the three stacked convolution layers, which restores the low-frequency features discarded by the attention mechanism and avoids the problem of gradient vanishing.

Based on the pixel-level shallow and high-level semantic feature maps (i.e., F_{DF} and F_0), we develop a palmprint image reconstruction subnetwork, which consists of one convolution layer and one PixelShuffle layer, for palmprint image SR as follows:

$$I_{\mathrm{SR}} = W_3 * H_{\mathrm{UP}}(F_{DF} + F_0) \tag{3}$$

where $I_{\mathrm{SR}} \in \mathbb{R}^{C_{\mathrm{in}} \times H \times W}$ is the final reconstructed SR palmprint image, $H_{\mathrm{UP}}(\cdot)$ denotes the learning of the PixelShuffle layer, and W_3 defines the weights of the 3×3 convolution layer at the end of the network.

B. Residual Hybrid Attention Block

To extensively exploit the multilevel semantic features of palmprints in the hybrid feature learning subnetwork, we develop multiple RHABs to simultaneously capture the global and local features of palmprint images. Fig. 2 shows the main body of the RHAB, which consists of double-stream branches, i.e., the CNN branch and transformer branch, for both local and global feature learning, respectively. Then, the local features and long-range dependencies captured from the two branches are progressively fused via feature fusion unit (FFU). In the following, we elaborate on the CNN branch, transformer branch, and FFU of the RHAB.

CNN Branch: In the RHAB, we develop a CNN branch to specifically learn the local features, such as line and edge characteristics of palmprint images. The CNN branch primarily consists of multiple enhanced hybrid attention modules (EHAMs) connected with embedded channel reduction and expansion layers. First, the reduction layer reduces the number of feature map channels whose ratio is set to 2, such that the memory storage and parameters of the network can be significantly reduced, and the expansion layer restores the number of channels to the original number, such that richer high-frequency features are conveyed in the local feature descriptor. The EHAM aims to adaptively emphasize the position-specific features of local regions via the attention mechanism. As such, the feature learning of the CNN branch can be expressed as

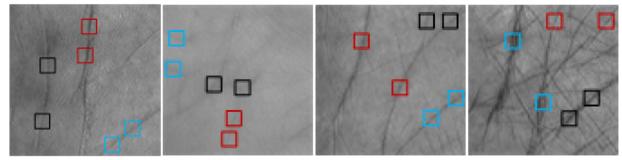


Fig. 3. Examples of the similar local patches in palmprint images with highly similar local characteristics, where the similar local patches are marked with the same color.

$$L_{\mathrm{local}} = H_{\exp}\left(f_{\mathrm{EHAM}}^{L}\left(f_{\mathrm{EHAM}}^{L-1}\left(\cdots f_{\mathrm{EHAM}}^{1}\left(H_{\mathrm{red}}\left(F_{\mathrm{in}}^{\mathrm{RHAB}}\right)\right)\right)\right)\right) \tag{4}$$

where $F_{\mathrm{in}}^{\mathrm{RHAB}}$ represents the input of the RHAB block. $f_{\mathrm{EHAM}}^{l}(\cdot)$ $(l = 1, 2, \ldots, L)$ indicates the feature mapping of the lth EHAM, $H_{\mathrm{red}}(\cdot)$ and $H_{\exp}(\cdot)$ represents the channel reduction and expansion layers, respectively, and L_{local} represents the local feature descriptor learned by the CNN-branch.

Transformer Branch: A palmprint image has many separated local patches with similar line or texture characteristics, as shown in Fig. 3, which can be used as a reference for palmprint image restoration, such as edge and texture patterns. Motivated by this, in the RHAB, we further develop a transformer branch to learn the long-range global information from similar patches based on the standard Swin transformer due to its strong feature expression ability of long-term dependency in the image.

Specifically, our proposed transformer branch consists of a reduction layer, several transformer blocks and an expansion layer. Similar to the CNN branch, the reduction and expansion layers are employed to reduce and recover the number of feature map channels, respectively. The transformer block adaptively learns the important position-specific global information to exploit the informative long-range dependencies and the global contextual features in the images. The learning process of the transformer branch can be expressed as

$$G_{\mathrm{global}} = H_{\exp}\left(f_{\mathrm{trans}}^{S}\left(f_{\mathrm{trans}}^{S-1}\left(\cdots f_{\mathrm{trans}}^{1}\left(H_{\mathrm{red}}\left(F_{\mathrm{in}}^{\mathrm{RHAB}}\right)\right)\right)\right)\right) \tag{5}$$

where G_{global} represents the global feature descriptor learned by the transformer-branch. $f_{\mathrm{trans}}^{s}(\cdot)$ $(s = 1, 2, \ldots, S)$ denotes the feature learning layer of the transformer block.

FFU: To combine the local and global features of palmprint images, we finally build an FFU by connecting a shared-weight multilayer perceptron (MLP) with a 3×3 convolution layer. First, the shared-weight MLP is employed to align the local and global feature representations within a shared semantic embedding space. By doing so, these two feature maps can be integrated in a more complementary way via a summation operation. Then, the 3×3 convolution layer further enhances the representation ability of the fused features. As a result, the

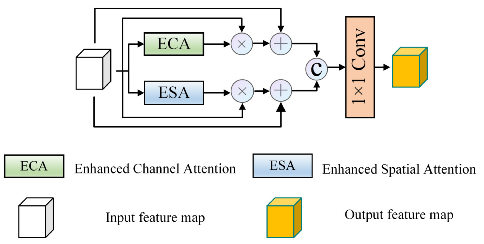


Fig. 4. Basic architecture of the EHAM includes an ECA for CA map learning, an ESA for spatial attention map generation, and a convolution layer for channel dimension reduction.

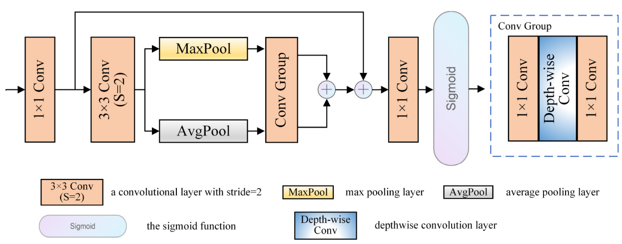


Fig. 5. Basic architecture of the ECA.

FFU fuses the local and global feature maps of the palmprint image as follows:

$$F_{\text{fuse}} = W_1^{\text{FFU}} * \left(H_{\text{mlp}}(L_{\text{local}}) + H_{\text{mlp}}(G_{\text{global}})\right) \tag{6}$$

where F_{fuse} denotes the final fused features. W_1^{FFU} represents the weights of 3×3 convolution layer at the end of FFU. $H_{\text{mlp}}(\cdot)$ is the mapping of MLP with one hidden layer. L_{local} and G_{global} represent the local and global feature maps, respectively.

C. Enhanced Hybrid Attention Module

To better capture the multitype local features of palmprint images, in the CNN branch of the RHAB, we design an EHAM by combining an ESA and an ECA for both spatial and CA palmprint feature map generation. Fig. 4 shows the basic architecture of the EHAM, including the ESA, ECA, and a 1×1 convolution layer for channel dimension reduction. In addition, multiple residual connections are utilized to stabilize the training.

ECA: Unlike the existing standard CA mechanism [39] that cannot fully exploit the interdependencies of feature channels with limited parameters, we propose an ECA module, as shown in Fig. 5, to completely exploit the interchannel relationship of palmprint feature maps.

For an input feature map of the ECA, e.g., $F_{\text{in}}^{\text{ECA}} \in \mathbb{R}^{C\times H\times W}$, ECA first uses a the 1×1 convolution layer to the number of input feature map channels and uses a 3×3 convolution layer with the stride $= 2$ to squeeze the spatial dimension, as follows:

$$F_1^{\text{ECA}} = W_2^{\text{ECA}} * W_1^{\text{ECA}} * F_{\text{in}}^{\text{ECA}} \tag{7}$$

where the W_1^{ECA} and W_2^{ECA} define the weights of the 1×1 and 3×3 convolution layers, respectively. $F_1^{\text{ECA}} \in \mathbb{R}^{C/r\times(H-1/2)\times(W-1/2)}$ refers to the convolutional result, where r is the reduction ratio. Then, ECA uses average pooling and max pooling to better-aggregate spatial information, as follows:

$$F_2^{\text{ECA}} = W_3^{\text{ECA}} * \text{AvgPool}\left(F_1^{\text{ECA}}\right) + W_3^{\text{ECA}} * \text{MaxPool}\left(F_1^{\text{ECA}}\right) \tag{8}$$

where AvgPool$(\cdot)$ and MaxPool$(\cdot)$ denote average pooling and max pooling, respectively. W_3^{ECA} represents the weights of the convolutional group, which contains two 1×1 convolution layers and a depthwise convolution layer. By using the shared-weight convolutional group (i.e., W_3^{ECA}), both average-pooled and max-pooled feature maps can be mapped to two CA maps under the same semantic space, which are further merged via a summation operation. Finally, a 1×1 convolution layer is employed to restore the number of channels of the merged CA maps (referred to as $F_3^{ECA} \in \mathbb{R}^{C\times 1\times 1}$). The feature learning of the ECA can be formulated as follows:

$$F_3^{\text{ECA}} = \sigma\left(W_4^{\text{ECA}} * \left(F_2^{ECA} + W_1 * F_{\text{in}}^{\text{ECA}}\right)\right) \tag{9}$$

where σ represents the sigmoid function. W_4^{ECA} denotes the final 1×1 convolution layer for feature channel restoration. It can be seen that the ECA focuses more on important channel feature learning of palmprint feature maps, such that more intrinsic characteristics, such as edges, textures, and lines of the palmprint, can be captured and used for final palmprint image SR.

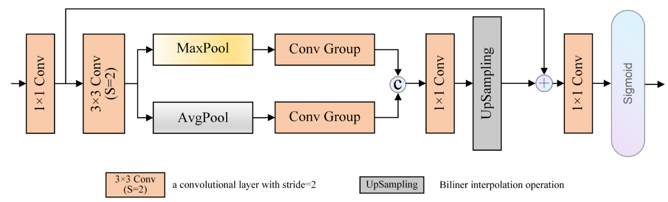


Fig. 6. Basic architecture of the ESA.

ESA: To further learn the important ROI from palmprint images, we introduce an ESA module based on the Convolutional block attention module [39] approach, as shown in Fig. 6.

Different from ECA, ESA first uses a 1×1 convolution layer to reduce channel information loss produced by pooling operation along the channel axis, and uses a 3×3 convolution layer to squeeze the spatial dimension of the feature map as follows:

$$F_1^{\text{ESA}} = W_2^{\text{ESA}} * W_1^{\text{ESA}} * F_{\text{in}}^{\text{ESA}} \tag{10}$$

where W_1^{ESA} and W_2^{ESA} are the weights of the 1×1 convolution layer and the 3×3 convolution layer with stride $= 2$, respectively. $F_1^{\text{ESA}} \in \mathbb{R}^{C/r\times(H-1/2)\times(W-1/2)}$ is the convolutional result. Then, the ESA extracts the spatial attention maps via average pooling and max pooling and improves the feature representation power via the convolution group as follows:

$$F_2^{\text{ESA}} = \text{Concat}\left(W_3^{\text{ESA}} * \text{AvgPool}\left(F_1^{\text{ESA}}\right), W_4^{\text{ESA}} * \text{MaxPool}\left(F_1^{\text{ESA}}\right)\right) \tag{11}$$

where W_1^{ESA} and W_2^{ESA} are the weights of the convolutional group. Furthermore, these two pooled spatial attention maps are fused (e.g., $F_{\text{fuse}}^{\text{ESA}} \in \mathbb{R}^{2\times(H-1/2)\times(W-1/2)}$) via concatenation, and the number of channels of the fused spatial attention map is reduced from 2 to 1 via a 1×1 convolution layer. It is noted that due to the convolution layer with stride = 2, the spatial dimensions of the feature maps are reduced. To address this, we utilize upsampling to restore the original spatial dimensions via bilinear interpolation. Finally, we use a 1×1 convolution layer to restore the number of channels of the upsampled feature maps from 1 to C.

The feature learning of the ESA can be summarized as follows:

$$F_3^{\text{ESA}} = \sigma\left(W_6^{\text{ESA}} * H_{\text{up}}\left(W_5^{\text{ESA}} * F_2^{\text{ESA}}\right) + W_1^{\text{ESA}} * F_{\text{in}}^{\text{ESA}}\right) \tag{12}$$

where W_5^{ESA} denotes the weights of the 1×1 convolution layer to reduce the number of channels. W_6^{ESA} is the weight of the 1×1 convolution layer to restore the number of channels. $H_{\text{up}}(\cdot)$ is the mapping of upsampling layer. $F_3^{\text{ESA}} \in \mathbb{R}^{C\times H\times W}$ denotes the spatial attention feature descriptor learned by the ESA.

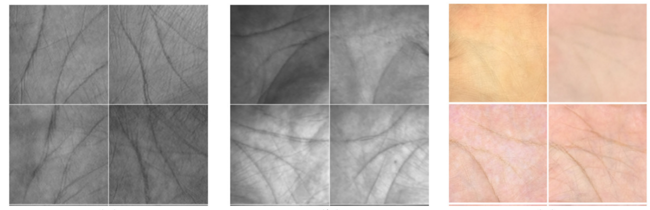
IV. Experiments

In this section, we conduct extensive comparative experiments on three widely used palmprint databases, including PolyU [4], TJI [5], and XJTU-UP [6] databases, to evaluate the proposed DHA network. For our proposed method, we set the number of EHAMs (i.e., L) and the number of transformer blocks (i.e., S) to 6. In the RHAB, we set the ratio r of the reduction and expansion layers to 2 for both the transformer and CNN branches and set the channel number (i.e., C) to 64 for the whole network. In addition, the ReLU [40] is used as the activation function of our network. Our proposed network is implemented on the PyTorch framework with using a single NVIDIA 3060Ti GPU, and it is trained for 150 epochs with a batch size of 32. The Adam optimizer with the hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.9999$ is employed for network training, and the initial learning rate is set to 2×10^{-4} with a reduce rate 2 for every 50 epochs. To compute the reconstructed palmprint image loss, we use the L_1 loss following the existing studies [14], [32], [36].

A. Datasets

The PolyU Multispectral Palmprint Dataset: The PolyU multispectral dataset contains four different illuminations, including blue, green, red, and near-infrared (NIR) illumination, of 24,000 palmprint images from 250 volunteers aged 20 to 60. Each subject was asked to capture 12 palmprint images under each spectrum. As a result, the PolyU dataset includes four different spectral palmprint image datasets. Each of the four spectral image databases contains 6,000 images captured from 500 palms of 250 individuals. In this experiment, we employ blue (PolyU-Blue) and NIR (PolyU-NIR) spectral palmprint images as the training datasets and use the green (PolyU-Green) and red (PolyU-Green) spectral palmprint images for image SR.

The TJI Palmprint Dataset: The TJI palmprint dataset consists of 12,000 images collected from 300 subjects. For an individual, each palm provided 20 palmprint images in two different sessions with an interval of approximately 61 days (10 images for each session). All palmprint samples were imaged in a contactless manner, leading to variant positions and poses of palms for different images.



(a) (b) (c)

Fig. 7. Some typical palmprint ROI images selected from the (a) PolyU, (b) TJI, and (c) XJTU-UP databases.

The XJTU-UP Dataset: The XJTU-UP dataset involves 20,000 palmprint images, which were captured from 100 subjects by using five different cell phones, including the iPhone 6S, HUAWEI Mate8, LG G4, Samsung Galaxy Note5, and MI8, in two different environments. Therefore, according to the types of the devices and the acquisition environments, the XJTU-UP dataset can be divided into ten different datasets, each of which contains 2,000 palmprint images from 200 different palms and 10 images for each palm.

In general, the original captured palmprint images usually contain nonpalm information, such as fingers and backgrounds. To address this, palmprint images are usually required to crop the ROI for palmprint image alignment and feature representation [4], [5], [6]. For this reason, in the experiment, we use the ROI of palmprint images for palmprint image SR. Fig. 7 shows the typical palmprint ROI samples provided by the PolyU, TJI and XJTU-UP databases. Moreover, in the experiments, each palmprint ROI image is randomly rotated with 90, 180 or 270 agrees, and randomly flipping in both vertical and horizontal directions to enhance the diversity of samples.

B. Palmprint Image SR

To evaluate the proposed method, we calculate the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) of the proposed DHA for palmprint image SR with different scale factors, including the $\times 2$, $\times 3$, and $\times 4$. Furthermore, we compare the proposed method against state-of-the-art methods, including the transformer-based methods, such as EDT [31], SwinIR [36], and SRFormer [41], and the typical CNN-based methods, such as EDSR [20], RCAN [13], second-order attention network (SAN) [33], HAN [34], and NLSA [35], and also the hybrid-based ACT [32] and HAT [27] method. For a fair comparison, all the compared methods were implemented under the same environment and were not pretrained. Table I tabulates the PSNRs and SSIMs obtained by different methods on the $\times 2$, $\times 3$, and $\times 4$ image SR tasks, where the top-two SR results are highlighted in red and blue, respectively.

Table I shows that our proposed method consistently outperforms the eight compared methods by achieving obviously better PSNRs and SSIMs. Compared with transformer-based

TABLE I
AVERAGE PSNR AND SSIM OBTAINED BY DIFFERENT METHODS FOR PALMPRINT IMAGE SR WITH THE SCALE FACTORS OF 2, 3, AND 4 ON THE POLYU-GREEN, POLYU-RED, TJI, AND XJTU-UP, RESPECTIVELY

Method	Scale	PolyU-Green		PolyU-Red		TJI		XJTU-UP	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [20]	×2	39.19	0.9113	41.74	0.9432	46.29	0.9809	39.19	0.9447
RCAN [13]	×2	39.31	0.9180	41.89	0.9450	46.37	0.9812	40.07	0.9501
SAN [33]	×2	39.34	0.9186	42.07	0.9452	46.51	0.9820	40.16	0.9567
HAN [34]	×2	39.32	0.9181	42.14	0.9461	46.54	0.9823	40.06	0.9548
NLSA [35]	×2	39.59	0.9217	42.17	0.9469	46.52	0.9821	40.18	0.9573
SwinIR [36]	×2	39.72	0.9230	42.24	0.9475	46.76	0.9826	40.25	0.9611
EDT [31]	×2	39.75	0.9232	42.27	0.9479	46.77	0.9827	40.28	0.9614
ACT [32]	×2	39.78	0.9237	42.29	0.9481	46.79	0.9831	40.32	0.9618
SRFormer [41]	×2	39.75	0.9234	42.28	0.9479	46.78	0.9829	40.31	0.9616
HAT [27]	×2	**39.79**	**0.9238**	**42.37**	**0.9485**	**46.83**	**0.9833**	**40.53**	**0.9629**
DHA(Ours)	×2	**39.85**	**0.9242**	**42.41**	**0.9494**	**46.91**	**0.9843**	**40.89**	**0.9634**
EDSR [20]	×3	36.86	0.8394	40.74	0.9212	41.24	0.9529	37.47	0.8824
RCAN [13]	×3	37.17	0.8496	40.81	0.9256	41.75	0.9560	37.51	0.8837
SAN [33]	×3	37.14	0.8487	40.79	0.9243	41.41	0.9532	37.50	0.8835
HAN [34]	×3	37.17	0.8497	40.82	0.9258	41.76	0.9561	37.53	0.8838
NLSA [35]	×3	37.28	0.8502	40.94	0.9262	41.78	0.9562	37.65	0.8843
SwinIR [36]	×3	37.84	0.8692	41.71	0.9403	41.82	0.9569	38.11	0.9007
EDT [31]	×3	37.82	0.8691	41.70	0.9402	42.14	0.9620	38.09	0.8997
ACT [32]	×3	37.88	0.8694	41.86	0.9411	42.27	0.9626	38.27	0.9125
SRFormer [41]	×3	37.86	0.8692	41.85	0.9410	42.25	0.9624	38.33	0.9137
HAT [27]	×3	**37.92**	**0.8699**	**41.94**	**0.9415**	**42.31**	**0.9629**	**38.46**	**0.9153**
DHA(Ours)	×3	**37.94**	**0.8712**	**42.05**	**0.9419**	**42.35**	**0.9635**	**38.55**	**0.9162**
EDSR [20]	×4	34.96	0.8315	38.52	0.9205	38.41	0.9311	37.17	0.8916
RCAN [13]	×4	35.36	0.8472	38.97	0.9120	38.76	0.9352	37.32	0.8923
SAN [33]	×4	35.42	0.8475	39.15	0.9129	38.83	0.9355	37.38	0.8929
HAN [34]	×4	35.42	0.8475	39.27	0.9135	38.91	0.9359	37.39	0.8929
NLSA [35]	×4	35.31	0.8465	39.28	0.9134	39.11	0.9378	37.42	0.8931
SwinIR [36]	×4	35.58	0.8481	39.83	0.9221	41.23	0.9385	37.68	0.8944
EDT [31]	×4	35.61	0.8479	40.09	0.9251	41.25	0.9386	37.71	0.8946
ACT [32]	×4	35.65	0.8483	40.11	0.9253	41.26	0.9387	37.72	0.8952
SRFormer [41]	×4	35.63	0.8481	40.10	0.9251	41.28	0.9389	37.70	0.8950
HAT [27]	×4	**35.68**	**0.8485**	**40.13**	**0.9255**	**41.31**	**0.9392**	**37.78**	**0.8957**
DHA(Ours)	×4	**35.73**	**0.8488**	**40.16**	**0.9261**	**41.32**	**0.9395**	**37.83**	**0.8967**

methods, such as EDT and SwinIR and SRFormer, our proposed DHA significantly enhances the quality of the reconstructed images by improving the PSNR by more than 0.04 dB in PSNR and the SSIM by 0.0006 for all scaling factors. This is because our proposed method develops an RHAB to collaboratively utilize and fuse the local and global features of palmprints, such that more intrinsic palmprint features can be exploited for context reconstruction of palmprint images. In contrast, existing transformer-based methods, such as SwinIR, usually directly use standard self-attention to capture global information while ignoring local information, making it difficult to capture fine-grained texture details of palmprints for HR image generation. In addition, our proposed method achieves higher PSNR and SSIM than CNN-based methods, such as EDSR and RCAN, with improvements of over 0.66 dB and 0.0109 across all scaling factors, respectively. The possible reason is that CNN-based methods, such as EDSR and RCAN, usually use limited receptive fields to capture local information and treat LR features equally across channel and spatial dimensions, making it difficult to recover various texture patterns of palmprint images. In contrast, our proposed method introduces a novel EHAM to enlarge the receptive field and adaptively consider interdependencies among channels and spatial dimensions to capture more context information, making the reconstructed palmprint images more faithful to the ground truth, such that better PSNR and SSIM can be obtained. In particular, the proposed DHA also outperforms the hybrid-based ACT and HAT in terms of the PSNR and SSIM. This is because unlike the ACT that simply combines the multitype features of samples at different learning stages and the HAT that combines the CA with self-attention in a parallel manner, our proposed DHA develops the ESA and ECA to specifically learn the small line and texture patterns of palmprints, such that high-perceptual quality of palmprint images can be obtained.

Fig. 8 shows the ×4 SR palmprint images obtained by different methods for a PolyU-Red sample. It is observed that the SR palmprint image generated by our proposed method shows more detailed structural and textural characteristics, such as edges and texture patterns with less noise corruption, while the other comparison approaches cannot well recover the rich texture characteristics of palmprint images and usually produce oversmoothness and oversharpness problems. This is because our proposed method designs the parallel learning branches to learn hierarchical palmprint features and the hybrid attention mechanism to further concentrate on small palmprint-specific pattern learning, such that clearer palmprint images can be generated.

C. Ablation Analysis

Our proposed DHA network embeds the ECA for exploiting the interchannel relationship of palmprint features, the ESA for utilizing the interspatial relationship of features and the FFU

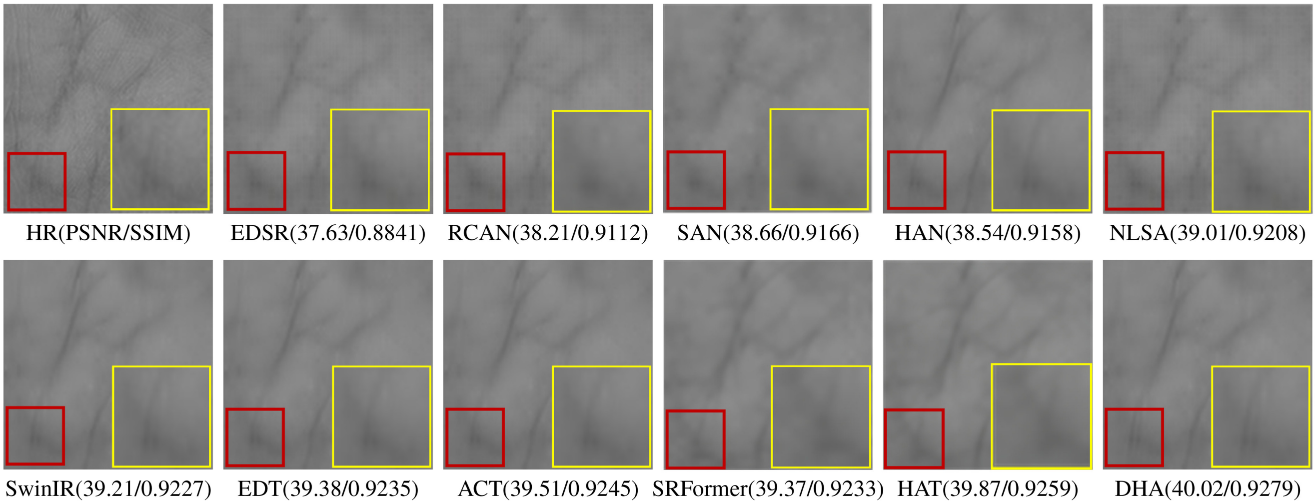


Fig. 8. SR palmprint images with a scale factor of obtained by different methods for a PolyU-red image.

TABLE II
PSNR AND SSIM OBTAINED BY OUR PROPOSED DHA FRAMEWORK WITH EMBEDDING CA AND ECA, RESPECTIVELY

Method	Scale	PolyU-Green		PolyU-Red		TJI		XJTU-UP	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DHA-CA	×2	39.80	0.9234	42.35	0.9490	46.68	0.9827	40.38	0.9612
DHA-ECA	×2	**39.85**	**0.9242**	**42.41**	**0.9494**	**46.91**	**0.9843**	**40.89**	**0.9634**
DHA-CA	×3	37.87	0.8691	41.86	0.9411	42.17	0.9628	38.25	0.9148
DHA-ECA	×3	**37.94**	**0.8712**	**42.05**	**0.9419**	**42.35**	**0.9635**	**38.55**	**0.9162**
DHA-CA	×4	35.67	0.8483	40.02	0.9251	41.17	0.9386	37.72	0.8959
DHA-ECA	×4	**35.73**	**0.8488**	**40.16**	**0.9261**	**41.32**	**0.9395**	**37.83**	**0.8967**

TABLE III
PSNR AND SSIM OBTAINED BY OUR PROPOSED DHA FRAMEWORK WITH EMBEDDING SA AND ESA, RESPECTIVELY

Method	Scale	PolyU-Green		PolyU-Red		TJI		XJTU-UP	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DHA-SA	×2	39.79	0.9232	42.08	0.9429	46.69	0.9811	40.31	0.9599
DHA-ESA	×2	**39.85**	**0.9242**	**42.31**	**0.9494**	**46.91**	**0.9843**	**40.89**	**0.9634**
DHA-SA	×3	37.86	0.8687	41.85	0.9409	42.21	0.9618	38.16	0.9125
DHA-ESA	×3	**37.94**	**0.8712**	**42.05**	**0.9419**	**42.35**	**0.9635**	**38.55**	**0.9162**
DHA-SA	×4	35.63	0.8473	40.01	0.9247	41.23	0.9375	37.56	0.8936
DHA-ESA	×4	**35.73**	**0.8488**	**40.16**	**0.9261**	**41.32**	**0.9395**	**37.83**	**0.8967**

for double-branch feature fusion. In the following, we conduct a series of ablation experiments to evaluate the effectiveness of these modules.

Analysis of ECA: To evaluate the ECA, we embed the conventional standard CA into our DHA framework and compare it with the proposed DHA with embedding the proposed ECA. Table II shows the PSNR and SSIM obtained by our DHA framework embedding the standard CA (DHA-CA) and the proposed ECA (DHA-ECA), respectively. It can be seen that our proposed method consistently achieves higher PSNR and SSIM than the DHA-CA on the four databases. This is because the standard CA mechanism directly pools the input features along the spatial axis to produce the CA map, usually making the spatial information loss significant. In contrast, our proposed ECA employs a strided convolution before pooling to aggregate more global features in a lower-spatial dimension, such that spatial information loss can be minimized.

Analysis of ESA: In the proposed DHA, we design an ESA to make the proposed network focus most on the feature learning of spatial ROI. To test the ESA, we embed the conventional standard SA into the DHA framework (DHA-SA) and compare it with the proposed DHA with embedding ESA (DHA-ESA). Table III tabulates the comparative results of the PSNR and SSIM obtained by the DHA-SA and DHA-ESA, respectively. We can see that our proposed method (i.e., DHA-ESA) outperforms the DHA-SA by obviously improving the PSNR and SSIM by approximately 0.01 dB and 0.002,

TABLE IV
PSNR AND SSIM OBTAINED BY OUR PROPOSED DHA FRAMEWORK WITH AND WITHOUT STACKED CNN LAYERS, RESPECTIVELY

Method	Scale	PolyU-Green PSNR	PolyU-Green SSIM	PolyU-Red PSNR	PolyU-Red SSIM	TJI PSNR	TJI SSIM	XJTU-UP PSNR	XJTU-UP SSIM
w/o stacked CNN	×2	39.81	0.9235	42.27	0.9479	46.85	0.9830	40.87	0.9631
DHA	×2	**39.85**	**0.9242**	**42.41**	**0.9494**	**46.91**	**0.9843**	**40.89**	**0.9634**
w/o stacked CNN	×3	37.92	0.8710	42.04	0.9418	42.28	0.9631	38.22	0.9142
DHA	×3	**37.94**	**0.8712**	**42.05**	**0.9419**	**42.35**	**0.9635**	**38.55**	**0.9162**
w/o stacked CNN	×4	35.69	0.8481	40.12	0.9258	41.25	0.9386	37.79	0.8963
DHA	×4	**35.73**	**0.8488**	**40.16**	**0.9261**	**41.32**	**0.9395**	**37.83**	**0.8967**

TABLE V
PSNR AND SSIM OF SINGLE-CNN BRANCH FRAMEWORK, SINGLE-TRANSFORMER BRANCH FRAMEWORK, AND PROPOSED DHA FRAMEWORK WITH EMBEDDING FFU AND WITH ELEMENTWISE ADDITION, RESPECTIVELY

Method	Scale	PolyU-Green PSNR	PolyU-Green SSIM	PolyU-Red PSNR	PolyU-Red SSIM	TJI PSNR	TJI SSIM	XJTU-UP PSNR	XJTU-UP SSIM
Single-CNN branch	×2	37.96	0.8967	40.11	0.9286	43.71	0.9543	38.13	0.9582
Single-transformer branch	×2	39.01	0.9168	41.96	0.9313	45.97	0.9793	39.57	0.9597
Fusion with add	×2	39.72	0.9217	42.22	0.9455	46.63	0.9827	40.36	0.9605
Fusion with FFU	×2	**39.85**	**0.9242**	**42.31**	**0.9484**	**46.91**	**0.9895**	**40.89**	**0.9634**
Single-CNN branch	×3	35.42	0.8274	38.96	0.8941	39.73	0.9341	37.64	0.9105
Single-transformer branch	×3	36.74	0.8437	40.67	0.9135	41.92	0.9586	38.01	0.9126
Fusion with add	×3	37.49	0.8593	41.79	0.9398	42.26	0.9623	38.21	0.9139
Fusion with FFU	×3	**37.94**	**0.8712**	**42.05**	**0.9419**	**42.35**	**0.9635**	**38.55**	**0.9162**
Single-CNN branch	×4	34.54	0.7941	36.76	0.8935	37.72	0.9005	36.54	0.8897
Single-transformer branch	×4	35.27	0.8104	38.97	0.9114	39.89	0.9341	37.42	0.8922
Fusion with add	×4	35.51	0.8449	40.12	0.9257	41.25	0.9358	37.65	0.8958
Fusion with FFU	×4	**35.73**	**0.8488**	**40.16**	**0.9261**	**41.32**	**0.9395**	**37.83**	**0.8967**

respectively. The possible reason is that the standard SA usually involves two pooling operations along the channel axis to merge channel information, leading to an increased channel information loss. Different from SA, our proposed ESA utilizes multiple 1×1 convolution layers to aggregate features from different channels, minimizing channel information loss. Furthermore, we utilize convolutional groups followed by two pooling operations instead of directly concatenating two feature maps in standard SA, which enhances the capability of spatial feature representation, such that higher PSNR and SSIM can be obtained.

Impact of Stacked CNN Layers: To evaluate the effectiveness of the "Stacked CNN Layers," we compare the performances of the proposed DHA between with embedding and without embedding the "Stacked CNN Layers," referred to as "w/o stacked CNN" and "DHA," respectively, and the ablation experimental results are tabulated in Table IV. We can see that the stacked CNN layers of the proposed method can obviously improve the SR performance over the DHA with no stacked CNN layers. This is because the stacked CNN layers can better compensate the loss of high-frequency information caused by self-attention, such that more texture and edge characteristics can be effectively exploited and higher PSNR and SSIM can be obtained.

Impact of Fusing Strategies: Our proposed DHA develops parallel CNN and transformer feature learning branches to extensively exploit the multitype features of palmprint images, which are then fused in a complementary way via the FFU. To evaluate the double feature learning branches and the FFU, we conduct the palmprint image SR experiments by using the single CNN-branch, single transformer-branch, double branch fusion via elementwise addition, and double branch fusion via the FFU (i.e., the proposed DHA), and Table V tabulates the PSNR and SSIM obtained by them. It can be seen that our proposed DHA consistently outperforms the single CNN-branch and single transformer-branch learning schemes. This is because the double branch structure can capture both the local and global information of palmprints, such that SR palmprint image reconstruction can be greatly benefited. In addition, our proposed method also achieves higher PSNR and SSIM than the double branch fusion scheme based on elementwise addition. The reason is that our proposed FFU employs a shared-weight MLP to align the local and global feature representations within a shared semantic embedding space. This alignment allows the two feature maps to be integrated in a more complementary way through the summation operation.

Mechanisms for Dense Connections: Unlike most existing forward blockwise transmission of feature maps, our propose DHA network uses dense connections to exploit more local features for palmprint image SR. To evaluate the effectiveness of the dense connection, we compare the performance of the proposed DHA with using and without using the dense connections (refer to as DHA-dense and DHA-straight), respectively, and the SR results of them are summarized in Table VI. It shows that the dense connection can obviously improve the SR performance of the DHA, demonstrating its effectiveness for retaining the shallow features to reconstruct more faithful palmprint images.

TABLE VI
PSNR AND SSIM OF PROPOSED DHA FRAMEWORK WITH EMBEDDING AND WITH NO DENSE CONNECTION, RESPECTIVELY

Method	Scale	PolyU-Green PSNR	PolyU-Green SSIM	PolyU-Red PSNR	PolyU-Red SSIM	TJI PSNR	TJI SSIM	XJTU-UP PSNR	XJTU-UP SSIM
DHA-straight	$\times 2$	39.81	0.9239	42.26	0.9481	46.88	0.9891	40.84	0.9628
DHA-dense		**39.85**	**0.9242**	**42.31**	**0.9484**	**46.91**	**0.9895**	**40.89**	**0.9634**
DHA-straight	$\times 3$	37.91	0.8703	42.01	0.9415	42.57	0.9666	39.17	0.9425
DHA-dense		**37.94**	**0.8712**	**42.05**	**0.9419**	**42.61**	**0.9668**	**39.25**	**0.9428**
DHA-straight	$\times 4$	35.93	0.8484	40.12	0.9259	41.21	0.9483	37.81	0.9207
DHA-dense		**35.97**	**0.8485**	**40.16**	**0.9261**	**41.27**	**0.9489**	**37.83**	**0.9209**

TABLE VII
IDENTIFICATION ACCURACY OF DIFFERENT METHODS ON THE POLYU_BLUE, POLYU_NIR, AND TJI DATASETS WITH SCALE FACTORS OF $\times 4$, RESPECTIVELY

Method	**TestSet**	**Train set size 1**	**2**	**3**	**4**	**5**
Compcode_LR/_SR	PolyU_Blue	87.02/94.37%	92.24/98.7%	96.63/99.44%	98.71/99.06%	97.06/99.38%
DON_LR/_SR		83.25/88.07%	90.12/93.57%	96.12/96.7%	96.8/98.94%	97.54/99.34%
ALDC_LR/_SR		91.65/97.49%	96.92/99.2%	98.32/99.86%	99.34/99.93%	99.66/99.94%
VGG_LR/_SR		42.69/45.47%	69.68/72.31%	77.18/85.78%	86.22/92.5%	88.34/96.51%
ResNet_LR/_SR		48.67/58.35%	69.34/81.94%	80.00/90.84%	86.08/94.65%	90.54/97.26%
Compcode_LR/_SR	PolyU_NIR	89.6/93.72%	96.23/98.18%	98.43/99.07%	98.8/99.34%	98.88/99.78%
DON_LR/_SR		82.96/89.16%	93.45/95.63%	93.65/98.11%	96.8/99.19%	98.64/99.39%
ALDC_LR/_SR		94.2/97.35%	98.62/99.02%	98.84/99.42%	99.21/99.67%	99.75/99.91%
VGG_LR/_SR		53.07/56.58%	76.18/79.08%	83.36/86.33%	88.42/90.26%	92.03/95.26%
ResNet_LR/_SR		43.71/48.33%	45.9/75.4%	78.13/88.2%	87.62/93.8%	91.14/96.11%
Compcode_LR/_SR	TJI	92.99/99.17%	97.31/99.79%	98.76/99.9%	99.02/99.91%	99.37/99.96%
DON_LR/_SR		93.11/95.46%	97.25/98.58%	98.66/99.29%	99.13/99.48%	99.18/99.61%
ALDC_LR/_SR		92.55/99.24%	97.18/99.82%	98.17/99.96%	98.75/99.93%	99.23/99.99%
VGG_LR/_SR		53.13/56.57%	76.54/82.14%	87.1/89.94%	89.56/92.43%	93.63/95.71%
ResNet_LR/_SR		64.13/66.97%	68.92/79.81%	82.05/91.38%	87.61/93.17%	91.23/96.57%

D. Palmprint Recognition Based on LR and SR Images

To further better evaluate the SR images obtained by our proposed SR method for palmprint recognition, we implement three representative palmprint descriptors, including the Compcode [42], ALDC [16], DON [43], and two typical feature learning networks, including the VGG [44] and ResNet [45] for palmprint recognition using the LR and SR palmprint images, respectively. We randomly select 1 to 5 palmprint images per palm as the training set and use the rest as the query set, and calculate the average rank-one identification accuracy. Table VII reports the rank-one identification rates of different methods on the PolyU-Blue, PolyU-NIR, and TJI datasets, where "_LR" and "_SR" denote the results of each method based on the original LR palmprint images and the SR palmprint images learned by our proposed DHA with the scale factor of $\times 4$, respectively. It can be seen that each method can obtain an obviously higher-identification accuracy on the SR images than that on the LR images. This demonstrates that our proposed method can better restore the intrinsic characteristics of palmprints, such as palmar texture patterns and various wrinkles, such that rich discriminative features can be provided for accurate palmprint recognition.

Furthermore, we implement diverse state-of-the-art SR methods, including EDT, ACT, SRFormer, and HAT, to generate the SR palmprint images of the TJI databases for palmprint recognition, and compare them with our proposed method. Table VIII presents the average rank-one identification accuracy rates of different methods based on the original LR palmprint images and the SR palmprint images generated by different SR methods. It can be seen that our proposed DHA can significantly improve palmprint identification accuracy when compared to the original LR palmprint images and the SR palmprint images generated by other SR methods. This is because our DHA can restore more local and global characteristics of palmprint images, such as edge patterns and texture information, from LR images using parallel CNN and transformer learning branches, effectively improving the performance of palmprint recognition methods.

TABLE VIII
IDENTIFICATION ACCURACY RATES OF DIFFERENT PALMPRINT RECOGNITION METHODS BASED ON THE ORIGINAL PALMPRINT IMAGES AND SR PALMPRINT IMAGES LEARNED BY DIFFERENT SR METHODS (WITH THE SCALE FACTORS OF $\times 4$ AND TRAINING SET SIZE OF 5 ON THE TJI DATASET

Method	**LR**	**EDT_SR**	**ACT_SR**	**SRFormer_SR**	**HAT_SR**	**DHA_SR**
Compcode	99.37%	99.84%	99.91%	99.89%	99.92%	**99.96%**
DON	99.18%	99.47%	99.55%	99.53%	99.56%	**99.61%**
ALDC	99.23%	99.95%	99.99%	99.97%	99.99%	**99.99%**
VGG	93.63%	95.55%	95.60%	95.58%	95.60%	**95.71%**
ResNet	91.23%	95.97%	96.42%	96.22%	96.43%	**96.57%**

E. Parameter Analysis

Fig. 9 shows the PSNR of our DHA network versus different number of the channel number of the 3×3 convolution

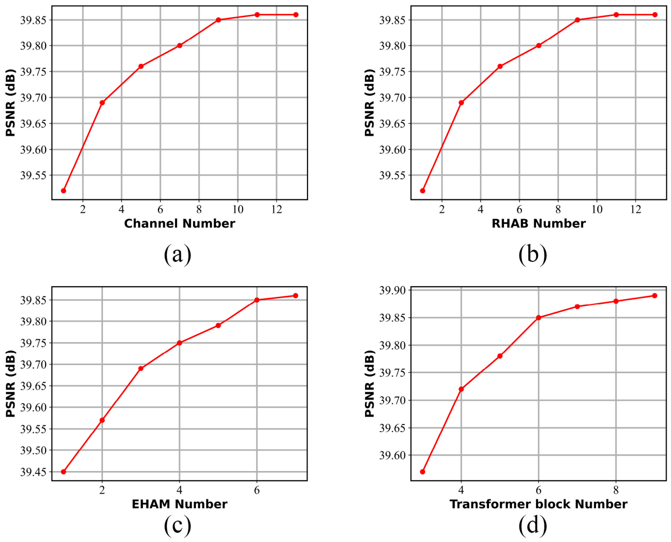


Fig. 9. PSNR of the proposed method versus (a) channel number, (b) RHAB number, (c) EHAM number, and (d) transformer block number, respectively, on PolyU-Green dataset for image SR (×2).

layer in shallow feature learning subnetwork, RHAB number, EHAM number in the CNN-branch and transformer block number in the transformer-branch, respectively. It is observed that the PSNR is obviously increased as the channel number increases from 32 to 64 and then the PSNR slightly fluctuates in about 80. This is because more diverse feature can be extracted via the multiple convolution layers as the channel number increasing until no more beneficial features. To balance the model size and also the image SR performance, we empirically set 64 as the channel number for our proposed network. For the RHAB number and EHAM number, we can see that the best PSNR can be obtained when the number of RHABs and EHAMs are 9 and 6, respectively, such that 9 RHABs and 6 EHAMs are empirically used for our proposed network. For the transformer block number, it can be seen that more transformer blocks generally improve the PSNR for image SR, because they allow the model to capture more complex long-dependency patterns and details. In this article, we empirically use 6 transformer blocks for the transformer-branch of our proposed network.

V. Conclusion

In this article, we propose a new DHA network for palmprint image SR. The proposed method first maps a palmprint image into a high-dimensional pixel-level representation and then simultaneously learns the local and global palmprint-specific features via parallel CNN-based and transformer-based learning branches. To clearly recover the intrinsic local palmprint patterns, we further design two enhanced spatial and CA modules to specifically capture the position-specific characteristics of palmprint images. Experimental results on six databases have demonstrated the promising effectiveness of our proposed DHA network for palmprint image SR.

References

[1] X. Liang, D. Zhang, G. Lu, Z. Guo, and N. Luo, "A novel multicamera system for high-speed touchless palm recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 3, pp. 1534–1548, Mar. 2019.

[2] S. Zhao and B. Zhang, "Learning salient and discriminative descriptor for palmprint feature extraction and identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5219–5230, Dec. 2020.

[3] A. Genovese, V. Piuri, K. N. Plataniotis, and F. Scotti, "PalmNet: Gabor-PCA convolutional networks for touchless palmprint recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, pp. 3160–3174, 2019.

[4] D. Zhang, Z. Guo, G. Lu, L. Zhang, and W. Zuo, "An online system of multispectral palmprint verification," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 2, pp. 480–490, Feb. 2009.

[5] L. Zhang, L. Li, A. Yang, Y. Shen, and M. Yang, "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach," *Pattern Recognit.*, vol. 69, pp. 199–212, Sep. 2017.

[6] H. Shao, D. Zhong, and X. Du, "Deep distillation hashing for unconstrained palmprint recognition," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, Jan. 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9335244

[7] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981.

[8] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Process.*, vol. 10, pp. 1521–1527, 2001.

[9] Y. Zhang, Q. Fan, F. Bao, Y. Liu, and C. Zhang, "Single-image super-resolution based on rational fractal interpolation," *IEEE Trans. Image Process.*, vol. 27, pp. 3782–3797, 2018.

[10] L. Wang, S. Xiang, G. Meng, H. Wu, and C. Pan, "Edge-directed single-image super-resolution via adaptive gradient magnitude self-interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1289–1299, Aug. 2013.

[11] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Trans. Image Process.*, vol. 21, pp. 4544–4556, 2012.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 184–199.

[13] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[14] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 12299–12310.

[15] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1874–1883.

[16] L. Fei, B. Zhang, Y. Xu, D. Huang, W. Jia, and J. Wen, "Local discriminant direction binary pattern for palmprint representation and recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 468–481, Feb. 2019.

[17] S. Zhao and B. Zhang, "Joint constrained least-square regression with deep convolutional feature for palmprint recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 1, pp. 511–522, Jan. 2020.

[18] L. Fei, B. Zhang, Y. Xu, C. Tian, I. Rida, and D. Zhang, "Jointly heterogeneous palmprint discriminant feature learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4979–4990, Sep. 2022.

[19] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2020.

[20] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 136–144.

[21] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2790–2798.

[22] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proc. 34th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3499–3509.

[23] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1646–1654.

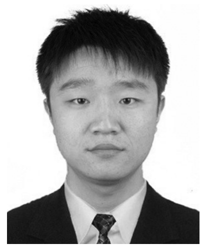
[24] Y. Tai, J. Yang, X. Liu, and C. Xu, “Memnet: A persistent memory network for image restoration,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4539–4547.
[25] C. Tian, Y. Xu, W. Zuo, C.-W. Lin, and D. Zhang, “Asymmetric CNN for image superresolution,” *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 6, pp. 3718–3730, Jun. 2021.
[26] J.-H. Kim, J.-H. Choi, M. Cheon, and J.-S. Lee, “Ram: Residual attention module for single image super-resolution,” 2018, *arXiv:1811.12043*.
[27] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, “Activating more pixels in image super-resolution transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 22367–22377.
[28] J. Liu, J. Tang, and G. Wu, “Residual feature distillation network for lightweight image super-resolution,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 41–55.
[29] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, “Residual non-local attention networks for image restoration,” 2019, *arXiv:1903.10082*.
[30] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, “Non-local recurrent network for image restoration,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–10.
[31] W. Li, X. Lu, J. Lu, X. Zhang, and J. Jia, “On efficient transformer based image pre-training for low-level vision,” 2021, *arXiv:2112.10175*.
[32] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim, “Enriched CNN-transformer feature aggregation networks for super-resolution,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 4956–4965.
[33] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 11057–11066.
[34] B. Niu et al., “Single image super-resolution via a holistic attention network,” in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 191–207.
[35] Y. Mei, Y. Fan, and Y. Zhou, “Image super-resolution with non-local sparse attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 3516–3525.
[36] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using swin transformer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2021, pp. 1833–1844.
[37] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9992–10002.
[38] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, “Early convolutions help transformers see better,” in *Proc. 35th Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 30392–30400.
[39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[40] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[41] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, “SRFormer: Permuted self-attention for single image super-resolution,” 2023, *arXiv:2303.09735*.
[42] A.-K. Kong and D. Zhang, “Competitive coding scheme for palmprint verification,” in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, pp. 520–523.
[43] Q. Zheng, A. Kumar, and G. Pan, “A 3-D feature descriptor recovered from a single 2-D palmprint image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1272–1279, Jun. 2016.
[44] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.
[45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

Yao Wang is currently pursuing the M.S. degree in artificial intelligence with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China.

His current research interests include image enhancement and biometrics.

Lunke Fei (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, Shenzhen, China, in 2016.

He is currently an Associate Professor with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. His research interests include pattern recognition, biometrics, image processing and machine learning.



Shuping Zhao (Member, IEEE) received the Ph.D. degree in computer science from the Department of Computer and Information Science, University of Macau, Macau, China, in 2020, and the M.S. degree in information science from South China Normal University, Guangzhou, China, in 2016.

He is currently with the School of Computer Science, Guangdong University of Technology, Guangzhou. He has published a number of technical papers at prestigious international journals, such as IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Systems, Man, and Cybernetics, and international conferences. His research interests include biometrics, machine learning, pattern recognition, and image processing.

Qi Zhu received the B.S., M.S., and Ph.D. degrees in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2007, 2010, and 2014, respectively.

He is a Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He has published over 100 scientific articles in refereed international journal and conference proceedings. His interests include pattern recognition, medical image analysis, and biometrics.

Jie Wen (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology (Shenzhen), Shenzhen, China, in 2019.

He is currently an Assistant Professor with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). He has authored or coauthored more than 60 technical papers at prestigious international journals and conferences, including the IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Image Processing, IEEE Transactions on Cybernetics, IEEE Transactions on Multimedia, PR, ECCV, AAAI, IJCAI, and ACM MM.

Dr. Wen serves/served as an Associate Editor for *International Journal of Image and Graphics* and the Area Chair for ACM MM. He served as the reviewer for more than 30 international journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, AAAI, IJCAI, CVPR, ICCV, and ACM MM. His research interests include image and video enhancement, pattern recognition, and machine learning.

Wei Jia (Member, IEEE) received the B.Sc. degree in informatics from Central China Normal University, Wuhan, China, in 1998, the M.Sc. degree in computer science from the Hefei University of Technology, Hefei, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence system from the University of Science and Technology of China, Hefei, in 2008.

He has been a Research Assistant and an Associate Professor with Hefei Institutes of Physical Science, Chinese Academy of Science, Hefei, from 2008 to 2016. He is currently a Research Associate Professor with the School of Computer and Information, Hefei University of Technology. His research interests include computer vision, biometrics, pattern recognition, image processing, and machine learning.

Imad Rida received the Ph.D. degree in computer science from Normandy University, Rouen, France, in 2017.

He is currently an Associate Professor with the University of Technologye of Compiègne, Compiègne, France. His research interests include machine learning, pattern recognition, and signal/image processing.