



OPEN

# Multi scale-aware attention for pyramid convolution network on finger vein recognition

Huijie Zhang, Weizhen Sun &amp; Ling Lv✉

In recent years, biometrics has been the most popular style of personal identification. The finger vein is an intrinsic and stable trait, and with the ability to detect liveness, it receives academic and industry attention. However, convolution neural networks (CNNs) based finger vein recognition generally can only cover a small input region by using small kernels. Hence, the performance is poor, facing low-quality finger vein images. It is a challenge to effectively use the critical feature of multi-scale for finger veins. In this article, we extract multi-scale features via pyramid convolution. We propose scale attention, namely, the scale-aware attention (SA) module, which enables dynamic adjustment of the weight of each scale to information aggregation. Utilize the complementation of different scale detail features to enhance the discriminativeness of extracted features, thus improving the finger vein recognition performance. In order to verify the present method's efficiency, we carried out experiments on two public data sets and one internal data, and the wide range of experimental results proves the proposed method's effectiveness.

Biometric identification is already widespread in everyday life, such as mobile payment, countries' banking systems, entrance guard systems, etc. Finger veins traits are a unique biometric trait via irradiating near-infrared (NIR) light, which is located beneath the skin tissue layer, and the deoxygenated hemoglobin in the blood of the finger veins absorbs near-infrared light energy. The finger vein non-contact acquisition can also bring a preponderance of finger vein traits: in vivo detection and high security. As the superiorities of finger vein traits, it is a promising approach by researchers. However, the advancement of finger vein recognition has fallen far short of expectations. The performance of finger vein recognition is susceptible to extrinsic and intrinsic factors, such as user behavior, ambient lighting, and light scattering within the finger tissues. As a result, for finger vein recognition and verification, it is urgent to design high-precision and robust finger vein feature extraction algorithms.

More generally accepted, feature extraction is the crucial step for automatic person recognition using finger vein, as verification performance is sensitive to it. In order to improve the performance, there are methods of image pre-processing applied to moderate the noise and enhance image contrast, such as adaptive histogram equalization (CLAHE) before feature extraction. Feature extraction can be classified into feature extraction-based hand-crafted and feature extraction-based deep learning. Despite the experimental results of hand-crafted being satisfactory, it inevitably got into elaborate pre-processing steps to reinforce vein texture. Besides this, the generalization ability of such methods is limited, owing to customized designs for some particular databases, and thus, they are only sometimes applicable to emerging databases. In recent years, convolutional neural networks have made tremendous progress in feature capabilities and remarkable image quality robustness, widely used in computer visions. Recently, with the growth of deep learning and the need for privacy of personal authentication, biometrics identification-based deep learning has been the trend, especially finger vein recognition, which has been the focus of scholars. Hou et al.<sup>1</sup> represent an auto-encoder based on Convolutional, which can learn high-level feature representation from raw finger vein images. Noh et al.<sup>2</sup> proposed that finger vein recognition integrates shape images with texture images based on a densely connected convolutional network, which can efficiently use diverse features and reduce the sensitivity to noise. Du et al.<sup>3</sup> proposed a method named FVSR-Net consisting of a multi-scale CNN and the improved biological optical model, which increases computational efficiency and stability. Fang et al.<sup>4</sup> proposed finger vein recognition based on a dual-stream convolutional network combined with incorporating both the mini-ROI and the original image, which achieved superior results. Wang et al.<sup>5</sup> presented a finger vein recognition approach with multi-receptive fields, which enhances the correlation of spaces and channels via dimensional interactive attention. Huang et al. designed an attention mechanism named

School of Biological Science & Medical Engineering, Southeast University, Nanjing 210096, China. ✉email: Lvling@njmu.edu.cn

joint attention (JA) to focus on the details of features, which dynamically adjusts the information aggregation in the spatial channels.

Inspired by these prior efforts and image detection, we presented an end-to-end backbone based on a novel mechanism for residual attention called “Scale-aware Attention (SA)”, which focuses on multi-receptive fields based on pyramid convolutional. The structure mainly consists of two stages: the feature extraction and the SA selection stages. In particular, it exploited pyramid convolutional to obtain different scale features from the input. This single-scale feature contains different details. Then, the SA choice stage leverages dynamic adjustment of the weights of different scale features to complement each feature. Finally, the refined feature was used for finger vein verification. The major contributions of the current work are outlined below:

- (1) We proposed a novel end-to-end model that replaces the stand convolutional with the pyramid convolutional. The proposed network uses a convolutional pyramid to extract multi-scale features via different kernel sizes to broaden receptive fields.
- (2) On the basis of SA, a coarse-to-fine feature extraction architecture is suggested which allows for end-to-end training, and the weights of different scale features can adjust adaptively. Using generalized mean pooling to generate the fused component weighted at different scales.
- (3) We conduct extensive experiments on public and in-house finger vein datasets, demonstrating proposed method outperforms existing SOTA methods while preserving good model and efficiency.

## Related work

### A brief review of pyramid convolution

Pyramid convolution is widely used in image classification since it uses different convolution kernels to expand the receptive field and complementary information of different scales. Wang et al.<sup>6</sup> present a multi-scale pyramid Convolution module with spatial attention and channel attention to pay more attention to the images’ object. Liang et al.<sup>7</sup> designed a model named SC2Net, which obtains the predicted density maps using residual pyramid dilated convolution (ResPyDConv). Jie et al.<sup>8</sup> deliver a well-traced spatial pyramid module that gathers global and local cues to solve the challenge of scale variation in deep convolutional networks. Jia et al.<sup>9</sup> combined pyramid-dilated convolutional blocks (PDCBs) with gated fusion units to reconstruct tiny image details since the PDCBs expand receptive fields and obtain details of the images via the network. Sun et al.<sup>10</sup> proposed a pyramid attention mechanism for classifying the Hyperspectral Images (HIS) based octave convolution network. Bao et al.<sup>11</sup> propose a residual attention pyramidal convolution for expression recognition-based VGG network, which uses CBAM attention to improve the weight of crucial aspects in multi-scale features extracted by pyramidal convolution. Moreover, it has better recognition performance and lightweight characteristics compared with other state-of-the-art networks. Due to the unique characteristics of finger vein images, like only grey channels and low contrast, extended receptive fields are a practical measure, and the studies on multi-receptive fields of finger veins mentioned in the previous section are successful example.

### Attention mechanism

The attention mechanism diverts more attention to essential regions and disregards irrelevant parts, imitating the human visual system. Moreover, it has exhibited excellent performance in various computer vision tasks, for instance, object detection, image classification, meta-learning, and individual recognition. Attention mechanisms can be categorized as temporal attention, spatial attention, channel attention, compound attention, and branch attention. SENet<sup>12</sup> is the pioneered attention channel, which, via a squeeze-and-excitation (SE) block, includes a squeeze module and excitation module to improve representation ability by adjusting the weight of the channel-wise relationship. Spatial transformer networks (STN)<sup>13</sup> are classical spatial attention systems that use a learnable module named spatial transformer to exploit spatial manipulation. Temporal attention is widely used for video processing to dynamically select when to pay attention. Selective kernel network (SKNet)<sup>14</sup> is classical branch attention, which uses soft attention to be guided by the information in these branches. The convolutional block attention module (CBAM)<sup>15</sup> represents compound attention, which recalibrates the significance of various spatial channels and positions by rescaling.

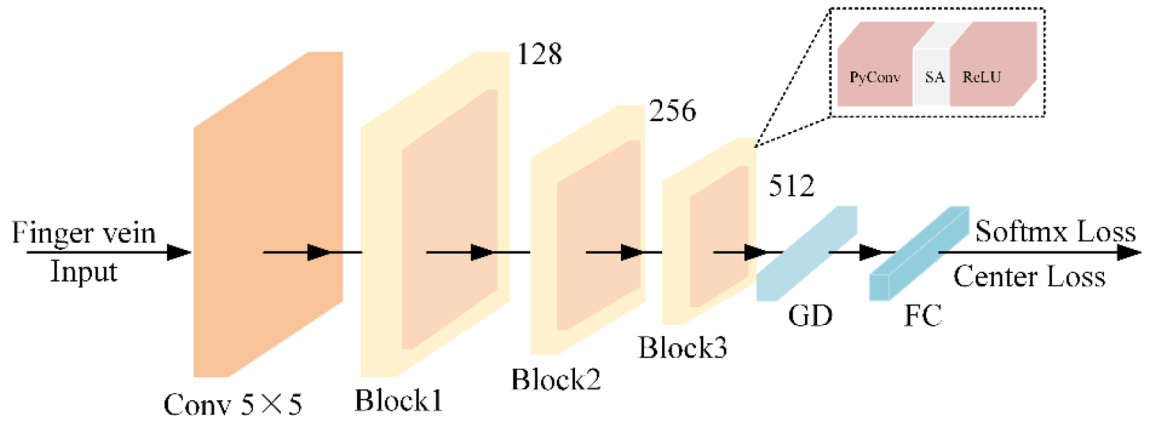
In finger vein authentication, there is much research about attention mechanisms. The attentions usually are a transformer of spatial attention, channel attention, and CBAM. As the multi-receptive field is extracted, the attention mechanism-based SKNet should be adequate for the multi-branch network, but it takes lots of computational power not applied. A new attention mechanism is designed from the scale level in this paper.

## Methodology

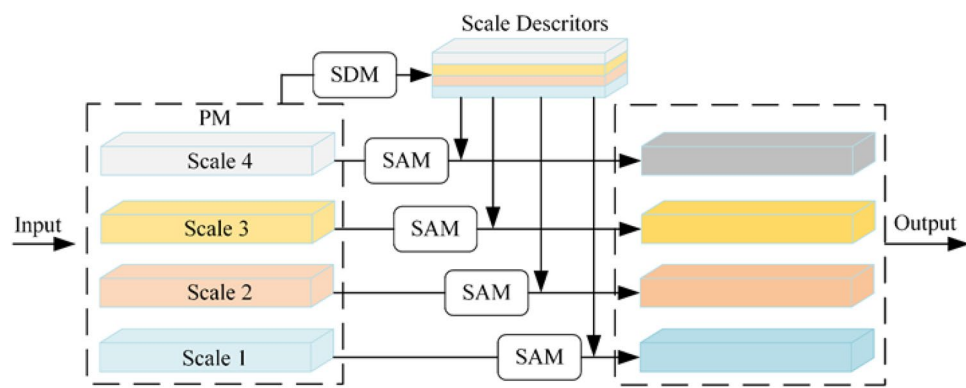
This section presents a new finger vein identification network with Residual scale-aware attention, namely RSAFVNet. The main task of RSAFVNet is to capture the details with a multi-scale, which captures the intense discrimination with an end-to-end structure. Figure 1 gives an overview of the structure of RSAFVNet. We designed a backbone named Residual scale-aware attention Pyramidal Convolution (RSAPyConv), which uses pyramidal convolution instead of strand convolution to extract multi-scale finger vein feature. The specific RASPyConv module block structure, as shown in Fig. 2, mainly includes the pyramidal convolution, scale-aware attention that consists of scale descriptor and scale attention block, the global descriptor, and the loss function.

### Pyramidal convolution

Convolutional neural networks are the workhorse for finger vein identification. The convolutional operation is the core of deep learning, which learns spatial kernels/filters for various computer vision tasks. Single kernel



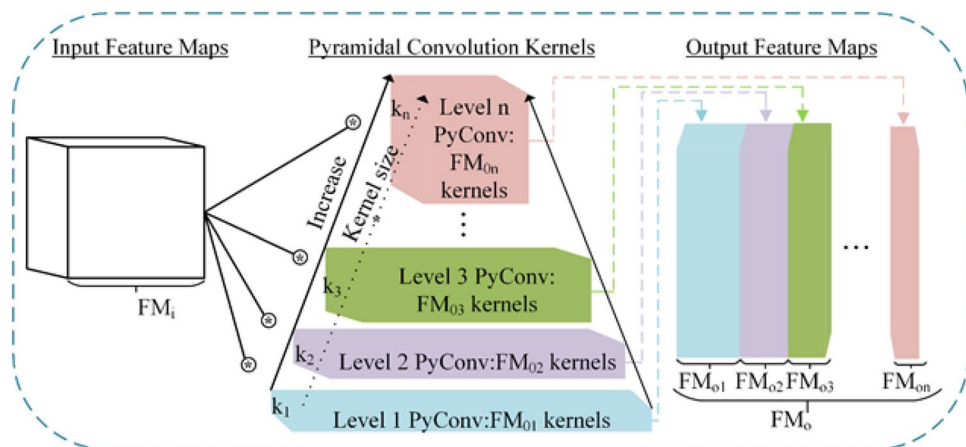
**Figure 1.** Overview of RSAFVNet.



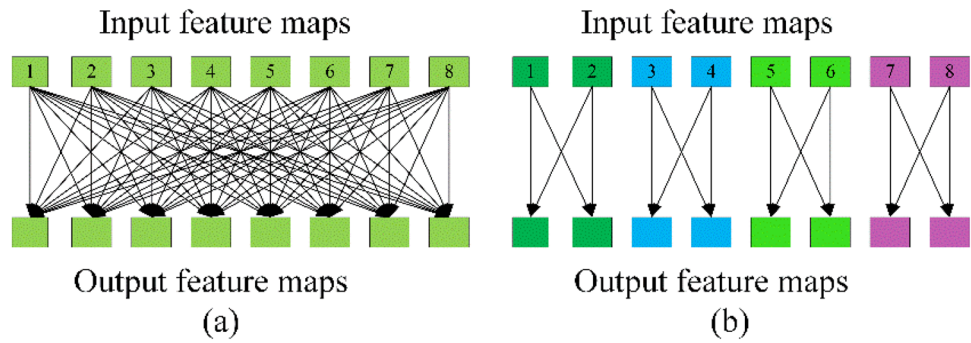
**Figure 2.** RASAPyConv's structure.

size standing convolution is used in most deep learning frameworks, which limits the network's receptive field. To expand the receiving field of the kernel, multi-branch CNNs were proposed, such as the representative InceptionNet families, but the amount of calculation was increased. Duta<sup>16</sup> proposed a multi-scale convolutional module named PyConv. An illustration is presented in Fig. 3, which displays a pyramid with  $n$  layers of various kernels, and the kernel has different spatial sizes for each level PyConv. With increasing the level of PyConv in the PyConv pyramid, the type of kernel size increased. In conjunction with the rise in space size, the depth of the kernel drops from level 1 to level  $n$ .

A varying kernel depth is applied at each level, and the input feature maps  $FM_i$  were separated into different groups. Two examples of various convolutions are displayed in Fig. 4. This example has eight output and input



**Figure 3.** Pyramid convolution (PyConv) structure.



**Figure 4.** Group convolution.

feature maps. The standard convolution is illustrated in Fig. 4a, where the depth of the kernel was equal to the number of input graphs. Consequently, each output feature map was associated with all input feature maps. Figure 4b indicates that the input feature maps were classified into four groups, each with the same kernel size. As a result, the kernel depth was decreased by four. Moreover, grouped convolution could significantly decrease the computational cost and the number of convolution parameters. On this basis, we introduced the PyConv into finger vein verification. Additionally, the critical issue for feature maps at various scales is how to capture the feature response between scales at each level.

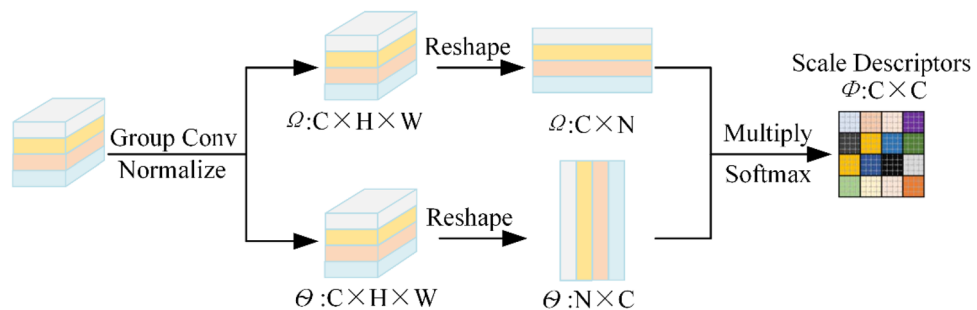
**Scale-aware attention**

To our knowledge, the finger vein images are remarkably similar, so the feature extracted for finger vein recognition requires high visual expression ability to address the slight difference in types of finger veins; multi-scale receptive fields were extracted in pyramidal convolution. Directly fusing the different scale maps to learn more discriminative features will drown out the valuable information in the redundant information volume. In light of the above analysis, we presented the SA to adaptively exploit complementary characteristics at various scales produced by the pyramidal convolution. The suggested SA comprises a scale attention module (SAM) and a scale descriptor module (SDM). SDM uses the different group maps to generate the scale descriptor. Meanwhile, the SAM combines the level descriptor to refine the scale features.

*Scale descriptor module*

The operational detail of SDM as illustration Fig. 5. The out of Pyramidal convolution includes multi-scale feature maps are fed into the SDM and the output is denoted as  $\Omega \in \mathbb{R}^{4C \times H \times W}$  and  $\Theta \in \mathbb{R}^{4C \times H \times W}$ . Firstly, SDM employs a  $1 \times 1$  convolutional layer to decrease the channel sizes. Then the SDM reshapes the  $\Omega$  to  $\mathbb{R}^{C \times N}$ , and  $\Theta \in \mathbb{R}^{N \times C}$  where  $N = H \times W$  represents the input finger vein image’s pixels number. And as the Pyramidal convolution uses four type of kernal sizes by groups,  $\Omega$  can be represented as  $\Omega = [\Omega_1, \Omega_2, \Omega_3, \Omega_4]$ ,  $\Theta$  can be represented as  $\Theta = [\Theta_1, \Theta_2, \Theta_3, \Theta_4]$ . The  $\Omega_i$  and  $\Theta_i \in \mathbb{R}^{\frac{C}{4} \times N}$  signal as the  $i$ -th scale. After that, SDM utilizes matrix multiplication between the transpose  $\Theta$  and  $\Omega$ , which is subsequently normalized with softmax lay to produce the scale descriptors:

$$\begin{aligned} \Phi &= \text{soft max}(\Omega \times \Theta) \\ &= \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_4 \end{bmatrix} = \begin{bmatrix} \Phi_{11} \cdots \Phi_{14} \\ \vdots \\ \Phi_{41} \cdots \Phi_{44} \end{bmatrix}, \end{aligned} \tag{1}$$



**Figure 5.** Detail of scale descriptor module.

where  $\Phi_{ij} = \Omega_i \Theta_j$ . As a result, the  $\Phi$  contains the information on the details of all scales. In this way, the scale descriptors produced by SDM contain both location and multi-scale information.

*Scale attention module*

We argue that single-scale features lack adequate semantic or spatial detail information to gain the performance of finger vein authentication. Hence, multi-scale information in the scale descriptors can augment single-scale features to give them greater discriminative power. In particular, as illustrated in Fig. 6, given the level descriptors  $\Phi$  and  $F_i, i \in [1, 2, 3, 4]$ , SAM adopts a matrix multiplication method to perform the perception of multi-scale information. Consequently, SAM acquires attentional scale characteristics with scale awareness. To further strengthen the feature representation, SAM applies a skip-connection approach here.

**Analysis performance of RASAPyConv**

RASAPyConv is the core of the design. According to the structure of Fig. 2, it can be known that during the feature extraction process, the input is first through the pyramidal convolution module, which generates multi-scale feature maps. Then, it is compressed by point convolution to bring the number of channels down to the number of single-scale channels, reducing network parameters. Finally, the feature maps through the SA attention mechanism are combined for dimension expansion. SA is mostly matrix transformation, so the central computation of the RASAPyConv module is generated by the PyConv module. The PyConv module is analyzed from two aspects: the number of parameters (space performance) and the computational requirements Floating Point Operations) (FLOPs) (time performance).

Assume that the PyConv module has  $FM_i$  as the number of input feature channels, the dimensions of each convolution layer are  $\{K_1^2, K_2^2, K_3^2, K_4^2\}$ , and the convolution depth is  $\left\{FM_i, FM_i \left(\frac{K_1^2}{K_2^2}\right), FM_i \left(\frac{K_1^2}{K_3^2}\right), FM_i \left(\frac{K_1^2}{K_4^2}\right)\right\}$  respectively, the number of feature channels output by convolution in each layer is  $\{FM_{o1}, FM_{o2}, FM_{o3}, FM_{o4}\}$  respectively, and the number of parameters together with FLOPs corresponding to PyConv module is as below:

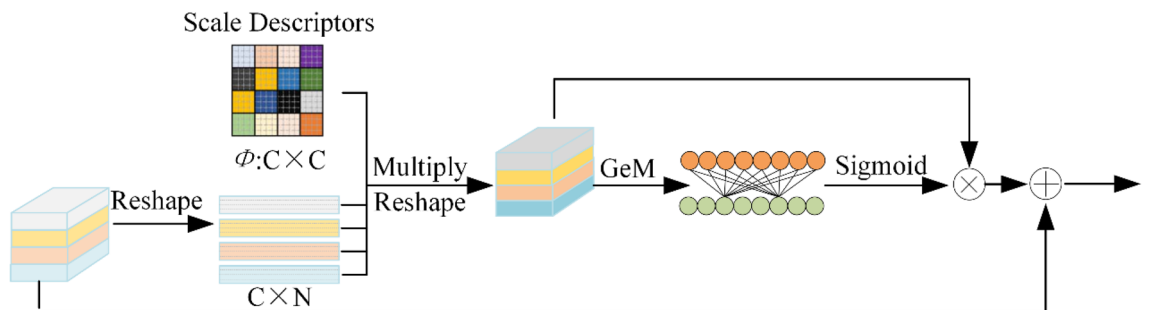
$$\begin{aligned}
 \text{parameters} = & K_1^2 FM_i FM_{o1} + K_2^2 FM_i \left(\frac{K_1^2}{K_2^2}\right) FM_{o2} \\
 & + K_3^2 FM_i \left(\frac{K_1^2}{K_3^2}\right) FM_{o3} + K_4^2 FM_i \left(\frac{K_1^2}{K_4^2}\right) FM_{o4},
 \end{aligned}
 \tag{2}$$

$$\begin{aligned}
 \text{FLOPs} = & K_1^2 FM_i FM_{o1} (H \times W) + K_2^2 FM_i \left(\frac{K_1^2}{K_2^2}\right) FM_{o2} (H \times W) \\
 & + K_3^2 FM_i \left(\frac{K_1^2}{K_3^2}\right) FM_{o3} (H \times W) + K_4^2 FM_i \left(\frac{K_1^2}{K_4^2}\right) FM_{o4} (H \times W).
 \end{aligned}
 \tag{3}$$

Among them, each of the four additive terms on the right-hand side of the equation corresponds to the number of parameters and the amount of computation for each layer of convolution in the PyConv module,  $\left(\frac{K_1^2}{K_i^2}\right)$  represents the number of groups of grouped convolutions of this convolutional layer,  $i = 1, 2, 3, 4$ .

Through formulation simplification, it can be found that although the size of the convolution kernel in the PyConv module increases from  $K_1^2$  to  $K_4^2$ , the parameters and calculation amount of each layer of convolution are equal to the parameters and calculation amount of the standard convolution of size  $K_1^2$ , which makes the PyConv module to expand the receptive domain of convolution kernel and improve the network performance without increasing the computational overhead. Therefore, compared with standard convolution, PM module has the following advantages.

- (1) *Multi-scale processing* The PyConv module has convolution kernels of different sizes and depths, which can parse input features from multiple scales and promote the full expression of finger vein features by fusing multi-scale features.



**Figure 6.** Details of scale attention module.



- (2) *High efficiency* The convolution of each layer of the PyConv module can realize independent parallel calculation, even run independently on different machines, and perform feature fusion. Hence, the overall calculation efficiency is high.

### Global feature descriptor

For biometrics, the authentication is to obtain a score that a feature vector extracted from the biometric traits' enrollment data compared with the obtaining. This is consistent with the image retrieval domain, which uses the global feature descriptor. Compared with the local features, it summarizes an image's content as a compact representation, and similarity learning is not limited by the object posed and the image's qualities. As those limitations also influence the performance of finger vein identification and the high similarity of local texture features, we employed global feature descriptors to produce a compact feature with multi-scale receptive fields for finger vein authentication.

GeM pooling can increase the features' representability via a nonlinear learnable operator, so it has been adopted for biometrics identification tasks<sup>17</sup>. Here, for a specific input finger vein image, the output of the RSAPyConv layer is a 3-dimensional tensor  $X$  with  $C \times W \times H$  size, indicating the number of channels, weight and height of feature maps separately. In computer vision, global max pooling (MAC<sup>18,19</sup>) and global average pooling (SPoC<sup>20</sup>) are generalized GeM pooling, which can be described as the below formulation:

$$f^{(g)} = [f_1^{(g)} \dots f_c^{(g)} \dots f_c^{(g)}], f_c^{(g)} = \left( \frac{1}{|X_C|} \sum_{x \in X_C} x^{p_c} \right)^{\frac{1}{p_c}}. \quad (4)$$

In which, the pooling parameter  $P_c$  is a learnable parameter, when  $p_c \rightarrow \infty$  or  $p_c = 1$ , the given form is the same as MAC or SPoC separately. The configurable and shareable pooling layer parameters allow for a highly non-linear presentation to be achieved with varying pooling parameters  $p_c$ . What is more, the GeM pooling operation is divisible and can be optimized via back-propagation, guaranteeing that our presented RSAFVNet can be trained end-to-end.

### Joint loss function strategy

The feature representation with small intra-class and significant inter-class distances can better promote finger vein recognition. In our experiment, a joint loss function, which combines softmax classification loss with a deep metric-based loss function (Center Loss function), was applied. Since the softmax loss only forces expand the inter-class distance of different classes, ignoring the intraclass distance, the finger veins' position and the lights may lead to a significant difference of the same finger vein in the capture, so that it may cause misclassification. The center loss function has a better ability to reduce the intraclass distance strongly, so we introduce it into the network, and the calculation process is as follows:

$$L_C = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|, \quad (5)$$

where  $x$  represents the input features,  $c_i$  denotes the center of all samples that have the same class label as  $y_i$ .

Therefore, the joint loss function calculation process is as follows:

$$L = L_s + \lambda L_c, \quad (6)$$

where  $\lambda$  is used to keep them in balance. And we will discuss its influence in the following experimental section.

## Experimental study

This section briefly describes the finger vein datasets utilized, the two public finger vein datasets, and an internal dataset. This section also describes the metrics that were evaluated. Then, it illustrates the detailed implementation, including default settings and training data preparation. Lastly, many experiments and their contents are employed to prove the effectiveness of the method presented, which includes ablation experiments, compared to other attention methods. At the same time, the superior performance of the presented method was proved compared to other state-of-the-art methods on a public benchmark dataset.

### Datasets and evaluation protocol

#### Datasets

- (a) *FV-USM (USM)* The FV-USM<sup>21</sup> dataset includes 2952 images in all, which was conducted via Universiti Sains Malaysia, George Town, Malaysia. Finger vein images were collected from the middle and index fingers of 123 people's hands, and they were gathered six times for each finger. The original images are grayscale and  $640 \times 480$  resolution. Also, there is a resolution of  $100 \times 300$  for the ROI provided in this dataset.
- (b) *SDUMLA-HMT (SDUMLA)* The SDUMLA<sup>22</sup> datasets includes 3816 images in all, which was conducted via Shandong University, Jinan, China. Finger vein images were collected from the ring, middle and index fingers of 106 people on both hands, and they were taken six times for each finger. The original images are grayscale and  $320 \times 240$  resolution.

- (c) *GERWIN* The *GERWIN*<sup>23</sup> datasets includes 8316 images in all, which was performed by our team. Finger vein images were collected from the ring, middle and index fingers of 106 people on both hands, and they were taken six times for each finger. The original images are grayscale and 600 × 200 resolution.

#### Evaluation protocol

In evaluation protocol, we adopt genuine pairs and impostor pairs, which introduced in<sup>24</sup>. Expressed mathematically as:

$$N_g = C_{N_f}^2 N_c = N_c N_f (N_f - 1) / 2, \quad (7)$$

$$N_i = C_{N_c}^2, \quad (8)$$

where the  $N_g$  represents the number of genuine pairs, the  $N_i$  represents the number of impostor pairs.  $N_c$  represents the number of finger vein categories in test set and  $N_f$  indicates the number of samples per finger vein category.

In order to fairly assess the behavior of various methods, equal error rate (EER) and accuracy (Acc) were calculated on a subset of tests. EER value equals the value when the false acceptance rate (FAR) equals the false rejection rate (FRR). The algorithms with a lower EER demonstrated improved performance on the validation tasks.

$$FAR = \frac{\text{Number of genuine rejection}}{\text{Number of genuine attempted}}, \quad (9)$$

$$FRR = \frac{\text{Number of imposter rejection}}{\text{Number of genuine attempted}}, \quad (10)$$

The open-set protocol was used in our experiment to assess the behavior of finger-vein verification, which is irrelevant to the actual application since the training classes and test classes do not overlap. In particular, for each database, one-half of the classes are randomized for training, and the remainder are applied for testing. In our trials, a fourfold cross-validation method was employed.

#### Implementation details

The above datasets have different resolutions, with redundant information and background fixed noise. We extracted the ROI image from those original images as the method applied in<sup>25</sup> and resized the finger vein samples to 224 × 224 × 3 for training. The grayscale one-channel sample was expanded to a three-channel image by copying twice as input. In<sup>26</sup> tradition, national data augmentation can improve the performance of finger veins and prevent overfitting. In our experiment, we adopt data augmentation methods such as flip and shear, etc., the specific ones listed in Table 1, which gives greater flexibility to train samples within each training cycle and to train finger vein samples without increase. The training samples of input are transformed randomly prior to entering each training epoch. In this format, the number of training samples remains constant.

In our experiments, the parameters are designed on the basis of previous experience of finger vein identification based on deep learning. The input batch size is 128, which includes 32 subjects and four samples in each subject. Furthermore, the epoch maximum is 100. The optimizer is an Adaptive Moment Estimation with an initial learning rate of 0.001 and dynamically adjusts using the ReduceLROnPlateau function. Specifically, the learning rate decreased to the original 0.1 when the validation set loss was not reduced within 20 batches. It should be pointed out that the experiment was conducted utilizing Python with the PyTorch frame on a workspace computer that had the below specifications: Intel(R)Core(TM) i7-8700 CPU @ 3.20 GHz RAM 16 GB, and GPU NVIDIA GeForce GTX 1060 6 GB.

#### Analysis of joint loss function strategy

##### Parameter selection

In order to assess the parameter  $\lambda$  performance of the network presented, in our tests, the scope of  $\lambda$  is between 0.01 and 5. The mean value was applied to determine the performance of the presented method, see Table 2.

Methods	Parameters	
Flip	Vertical	True
	Horizontal	False
Rotation (degree range for random rotation)		2.5
Shift (range scale in [-1, 1])	Width	0.05
	Height	0.05
Shear range (range in [-1,1])		0.05
Zoom (range in [1 - zoom, 1 + zoom range])		0.05

**Table 1.** Data augmentation parameters.

Table 2 displays the performance of the presented method in various datasets. This indicates that the suitability of the parameter  $\lambda$  can improve the recognition of finger veins. For instance, in the FVUSM dataset, the optimum performance of the presented method is achieved with  $\lambda$  equals to 0.1, the Acc of 99.92%, and EER of 1.02%. Similarly, in the SDU datasets, the optimum performance of the presented method with the  $\lambda$  equal 0.5, the Acc of 98.95%, and the EER of 0.73%. As for the GERWIN dataset, the proposed method's best performance with the  $\lambda$  equals 2, the Acc of 99.16%, and EER of 0.23%.

#### Loss function comparison

To assess the effectiveness of the joint loss function strategy, softmax, and the central loss function were compared to the joint loss function. The performance of different loss functions is listed in Table 3 in detail.

As illustrated in Table 3, the properties of the joint loss function-based FV-USM dataset are compared to the softmax loss and central loss functions; the Acc increased by 1.74% and 0.52%, and the EER increased by 1.33% and 0.40%, respectively. The performance of the SDUMLA-HMT datasets based on the joint loss function compared to the softmax loss and central loss functions, the Acc increased by 3.12% and 0.93%, and the ERR increased by 3.56% and 0.58%, respectively. The performance of the GERWIN datasets based on the joint loss function compared to the softmax loss and central loss functions, the Acc increased by 1.00% and 0.29%, and the ERR increased by 3.52% and 1.45%, respectively.

#### Effectiveness of the pyramidal convolution

An experiment was performed on SDUMLA-HMT, FV-USM, and GERWIN datasets to validate the significance of the pyramidal convolution. The pyramidal convolution was replaced by standard convolution, which has the same number of channels. The output of the standard convolution reduces the channel to one in four and splits it into four groups, achieving the scale-aware attention mechanism with a single scale. The performance of different convolutions is listed in Table 4. As shown, the pyramidal convolution outperforms the standard convolution.

#### Effectiveness of the SA attention module

##### Ablation study of SA attention

In order to investigate the properties of the proposed module of SA, an ablation study of SA was performed on datasets of SDUMLA-HMT, FV-USM, and GERWIN. Furthermore, SA attention evolved from PA attention, contrasting with PA attention. The findings of the SA ablation study are presented in Table 5.

As shown in Table 5, the performance of the FV-USM datasets with SA attention compared with no attention and PA attention increased by 0.89% and 0.46%, and EER increased by 0.52% and 0.19%, respectively. The performance of the SDUMLA-HMT datasets with SA attention compared with no attention and PA attention, the Acc increased by 1.82% and 0.08%, and ERR increased by 0.64% and 0.39%, respectively. Compared with no attention and PA attention, the performance of the GERWIN datasets with SA attention increased by 1.62% and 0.47%, and ERR increased by 1.13% and 0.17%, respectively.

##### Compared with different attention

To better demonstrate the effectiveness of the presented scheme, two classical attentional mechanisms, CBAM and SE, were compared with the newly presented finger-vein joint attention. The optimum values are represented

$\lambda$	FV-USM		SDUMLA-HMT		GERWIN	
	Acc (%)	EER (%)	Acc (%)	EER (%)	Acc (%)	EER (%)
0.01	99.86	1.35	96.38	4.92	97.56	3.65
0.05	99.89	1.22	97.02	2.31	98.87	2.86
0.1	99.92	1.02	98.70	1.30	99.10	1.32
0.5	99.91	1.06	98.95	0.73	99.05	0.65
1	99.91	1.10	98.15	0.95	99.08	0.36
2	99.90	1.12	97.92	1.45	99.16	0.23
5	99.90	1.12	97.95	1.64	98.52	2.32

**Table 2.** Performance of different loss functions in different datasets.

Loss function	FV-USM		SDUMLA-HMT		GERWIN	
	Acc (%)	EER (%)	Acc (%)	EER (%)	Acc (%)	EER (%)
Softmax loss	98.18	2.35	96.83	4.29	98.16	3.65
Center loss <sup>27</sup>	99.40	1.42	98.02	1.31	98.87	1.68
The proposed loss	99.92	1.02	98.95	0.73	99.16	0.23

**Table 3.** Performance comparison of different loss functions.



Convolution	FV-USM		SDUMLA-HMT		GERWIN	
	Acc (%)	EER (%)	Acc (%)	EER (%)	Acc (%)	EER (%)
Standard	99.35	1.32	87.76	3.73	89.34	6.56
Pyramidal <sup>16</sup>	99.92	1.02	98.95	0.73	99.16	0.23

**Table 4.** Performance comparison of different convolutions.

Setting	FV-USM		SDUMLA-HMT		GERWIN	
	Acc (%)	EER (%)	Acc (%)	EER (%)	Acc (%)	EER (%)
No attention	99.03	1.54	97.67	1.37	97.54	2.36
+ PA	99.46	1.21	98.87	1.12	98.69	0.40
+ SA(ours)	99.92	1.02	98.95	0.73	99.16	0.23

**Table 5.** Performance of SA.

in bold, as illustrated in Table 6. From these results, each attention module outperforms the baseline modules with no attention. Nevertheless, the module with SA attention presents the best performance on finger vein public datasets and in-house datasets with the highest Acc and the lowest EER. These present the validity of the method proposed. Figure 7 displays the findings of the two former experiments, where performance is better as the curve gets closer to the axis, and the intersection point between the diagonal and the curve is the EER.

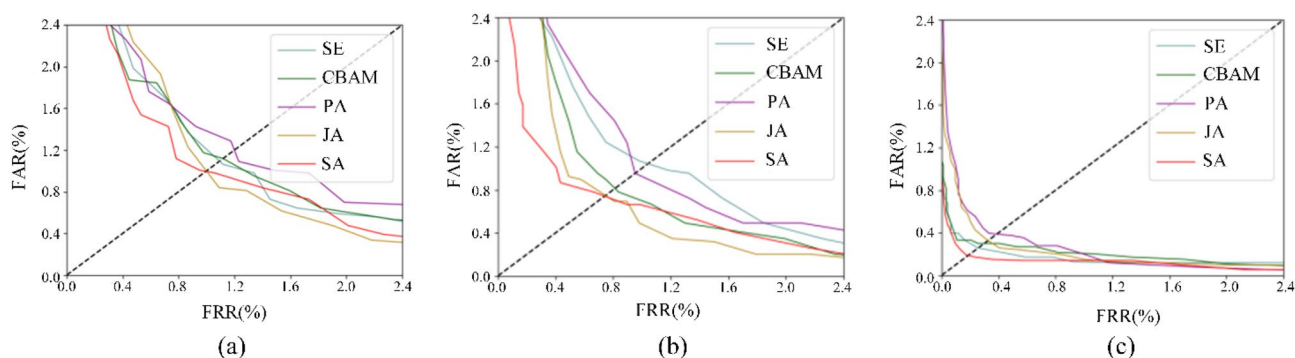
### Comparison with SOTA methods

We assessed our presented network through a comparison with state-of-the-art finger vein authentication approaches. The tests were performed on the public datasets SDUMLA and FV-USM. The findings are displayed in Tables 7 and 8, and the bold indicates the optimum value.

As we can see from the tables, our method has a slight advantage over the method proposed in Ref.<sup>28</sup>, and the method of extracted feature proposed does not need prior knowledge. Fang et al. proposed a double-weighted group sparse representation classification with some finger vein details lost, while our proposed method to extract multi-scale finger vein details feature. Liu et al.<sup>31</sup> proposed attention by connecting a residual block with multistage residual attention. They achieved a good performance in public datasets, and Huang et al.<sup>32</sup> presented a new transformer based on finger vein recognition, focusing on global information. However, the proposed

Setting	FV-USM		SDUMLA-HMT		GERWIN	
	Acc (%)	EER (%)	Acc (%)	EER (%)	Acc (%)	EER (%)
No attention	99.03	1.54	97.67	1.37	97.54	2.36
+ SE	98.14	1.12	98.12	1.15	98.35	0.28
+ CBAM	98.56	1.15	95.65	0.81	99.03	0.30
+ JA <sup>17</sup>	99.67	1.04	98.92	0.75	99.12	0.31
+ SA(ours)	<b>99.92</b>	<b>1.02</b>	<b>98.95</b>	<b>0.73</b>	<b>99.16</b>	<b>0.23</b>

**Table 6.** Comparisons of different attention modules.



**Figure 7.** Comparison of EER for attention modules in various databases. (a) SDUMLA-HMT. (b) FV-USM. (c) GERWIN.

Paper	Method	SDUMLA-HMT	
		Acc (%)	EER (%)
Li et al. <sup>28</sup>	Compact multirepresentation feature descriptor	96.43 ± 2.45	–
Fang et al. <sup>29</sup>	Double-weighted group Spare representation	70.56	–
Dev et al. <sup>30</sup>	Multi Biometric finger vein template	99.03	3.87
Liu et al. <sup>31</sup>	Finger vein recognition based MMRAN	98.11	0.44
Huang et al. <sup>32</sup>	FVT	97.90	1.50
Vida et al. <sup>33</sup>	Cuboc-LBP	98.50	1.35
Ours	Scale-aware attention network	<b>98.95</b>	<b>0.73</b>

**Table 7.** Comparisons with the state-of-art on SDUMLA-HMT.

Paper	Method	FV-USM	
		Acc (%)	EER (%)
Li et al. <sup>28</sup>	Compact multirepresentation feature descriptor	97.69 ± 2.95	–
Fang et al. <sup>29</sup>	Double-weighted group Spare representation	97.00	–
Liu et al. <sup>31</sup>	Finger vein recognition based MMRAN	99.75	0.50
Vida et al. <sup>33</sup>	Cuboc-LBP	99.50	0.95
Huang et al. <sup>32</sup>	FVT	99.73	0.44
Ours	Scale-aware attention network	<b>99.92</b>	<b>1.02</b>

**Table 8.** Comparisons with the state-of-art on FV-USM.

method is superior to MMRAN and FVT, and the design of the finger vein model of the end-to-end multi-scale attention mechanism in this paper is also inspired by them. Overall, the proposed method has the lowest EER and the highest Acc of the compared methods, which indicates that the proposed method is effective.

## Conclusion

This paper presented a novel end-to-end finger vein authentication network based on the pyramidal convolution with a scale-aware attention module block. Pyramid convolution learned finger vein features at different scales, and the scale-aware captured various detailed features with multi-scale, avoiding the redundancy of finger vein features. By jointly applying the central loss and softmax loss functions, the inter-class clustering was increased, and the distance of the intraclass was reduced. The model's effectiveness was verified on the open-source finger vein dataset and our in-house dataset. The ablation experiment demonstrated the effectiveness of the pyramid convolution module, joint loss function stagey, and the scale-aware attention mechanism module on recognizing the finger vein.

Although the proposed method performs significantly in finger vein authentication, some issues still need further research. Firstly, the finger vein collected will be mixed with noise in the collection process to extract more compelling features and prevent illegal people from introducing potential backdoor attack threats. Second, different types of finger vein data are collected in different usage scenarios, which requires the model to have high generalization performance. Still, future work can consider cross-validation experiments to contribute to the optimization of model parameters and improve the generalization capability. Finally, with the increase of users, the query authentication time will also increase significantly, so more convenient authentication methods are also a focus of future research.

## Data availability

The data that support the findings of this study are available from SDUMLA (<https://time.sdu.edu.cn/kycg/gksjk.htm>) and FV-USM ([http://drfendi.com/fv\\_usm\\_database/](http://drfendi.com/fv_usm_database/)) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding author upon reasonable request and with permission of official. GERWIN dataset generated and analysed during the current study are not publicly available due the requirements of all the subjects but are available from the corresponding author on reasonable request.

Received: 21 July 2023; Accepted: 28 December 2023

Published online: 04 January 2024

## References

- Hou, B. R. & Yan, R. Q. Convolutional auto-encoder model for finger-vein verification. *IEEE Trans. Instrum. Meas.* **69**(5), 2067–2074 (2019).
- Noh, K. J., Choi, J., Hong, J. S. & Park, K. R. Finger-vein recognition based on densely connected convolutional network using score-level fusion with shape and texture images. *IEEE Access* **8**, 96748–96766 (2020).

3. Du, S. S., Yang, J. F., Zhang, H. G., Zhang, B. & Su, Z. G. FVSR-Net: An end-to-end finger vein image scattering removal network. *Multimedia Tools Appl.* **80**(7), 10705–10722 (2021).
4. Fang, Y. X., Wu, Q. X. & Kang, W. X. A novel finger vein verification system based on two-stream convolutional network learning. *Neurocomputing* **290**, 100–107 (2018).
5. Wang, K. X., Chen, G. H. & Chu, H. J. Finger vein recognition based on multi-receptive field bilinear convolutional neural network. *IEEE Signal Process. Lett.* **28**, 1590–1594 (2021).
6. Wang, G. H., Cheng, L., Lin, J. H., Dai, Y. Y. & Zhang, T. L. Fine-grained classification based on multi-scale pyramid convolution networks. *PLoS ONE* **16**(7), e0254054 (2021).
7. Liang, L. J. *et al.* SC2Net: Scale-aware crowd counting network with pyramid dilated convolution. *Appl. Intell.* **53**(5), 5146–5159 (2022).
8. Jie, F. R., Nie, Q. F., Li, M. S., Yin, M. & Jin, T. S. Atrous spatial pyramid convolution for object detection with encoder–decoder. *Neurocomputing* **464**, 107–118 (2021).
9. Jia, X. L. *et al.* Pyramid dilated convolution neural network for image denoising. *J. Electron. Imaging* **31**(2), 023024 (2022).
10. Sun, Y. F. *et al.* Resolution reconstruction classification: Fully octave convolution network with pyramid attention mechanism for hyperspectral classification. *Int. J. Remote Sens.* **43**(6), 2076–2105 (2022).
11. Bao, Z. & Chen, H. H. Expression recognition based on residual attention mechanism and pyramid convolution. *Pattern Recogn. Artif. Intell.* **35**(6), 497–506 (2022).
12. Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. H. Squeeze-and-excitation network. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(8), 2011–2013 (2020).
13. Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuolu, K. Spatial transformer networks In *Proc. 28th International Conference on Neural Information Processing Systems*, Vol. 2, 2017–2025 (2015).
14. Li, X., Wang, W. H., Hu, X. L. & Yang, J. Selective kernel networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 510–519 (2019).
15. Woo, S., Park J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *Proc. Eur. Conf. Comput. Vis. (ECCV)* 3–19 (2018).
16. Duta, I. C., Liu, L., Zhu, F. & Shao, L. Pyramidal convolutional: Rethinking convolutional neural networks for visual recognition. <http://arXiv.org/2006.11538>.
17. Huang, J. D., To, M., Yang, W. L. & Kang, W. X. Joint attention network for finger vein authentication. *IEEE Trans. Instrum. Meas.* **70**, 1–11 (2021).
18. Toliás, G., Sicre, R. & Jégou, H. Particular object retrieval with integral max-pooling of CNN activations. <http://arxiv.org/abs/1511.05879> (2015).
19. Razavian, A. S., Sullivan, J., Maki, A. & Carlsson, S. Visual instance retrieval with deep convolutional networks. *ITE Trans. Media Technol. Appl.* **4**(3), 251–258 (2016).
20. Babenko, A. & Lempitsky, V. Aggregating deep convolutional features for image retrieval. <http://arxiv.org/abs/1510.07493> (2015).
21. Asaari, M. S. M., Suandi, S. A. & Rosdi, B. A. Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics. *Expert Syst. Appl.* **41**(7), 3367–3382 (2014).
22. Yin, Y., Liu, L. & Sun, X. SDUMLA-HMT: A multimodal biometric database. In *Biometric Recognition (Lecture Notes in Computer Science)* Vol. 7098 (eds Sun, Z. *et al.*) (Springer, 2011).
23. Hou, B. R., Zhang, H. J. & Yan, R. Q. Finger-vein biometric recognition: A review. *IEEE Trans. Instrum. Meas.* **71**, 1 (2022).
24. Kauba, C. & Uhl, A. An available open-source vein recognition framework. In *Handbook of Vascular Biometrics* (eds Uhl, A. *et al.*) 113–142 (Springer, 2020).
25. Gupta, P. & Gupta, P. An accurate finger vein based verification system. *Digit. Signal Process.* **38**, 43–52 (2015).
26. Ou, W. F., Po, L. M., Zhou, C., Rehman, Y. A. U. & Xian, P. F. Fusion loss and inter-class data augmentation for finger vein feature learning. *Expert Syst. Appl.* **171**, 114584 (2021).
27. Wen, Y., Zhang, K., Li, L. & Qiao, Y. A discriminative feature learning approach for deep face recognition. In *Proc. Eur. Conf. Comput. Vis.*, Vol. 9911, 499–515 (Springer, 2016).
28. Li, S. Y., Ma, R. J., Ma, L. K. & Zhang, B. Learning compact multirepresentation feature descriptor for finger-vein recognition. *IEEE Trans. Forensics Secur.* **17**, 1946–1958 (2022).
29. Fang, C. X., Ma, H., Yang, Z. D. & Tian, W. B. A finger-vein recognition method based on double-weighted group sparse representation classification. *Int. J. Mach. Learn. Cybern.* **13**, 2725–2744 (2022).
30. Dev, R., Tripathi, R. & Khanam, R. Q. Y. A novel approach based multi biometric finger vein template recognition system using HGF. *Open Comput. Sci.* **11**, 337–345 (2021).
31. Liu, W. Y. *et al.* MMRAN: A novel model for finger vein recognition based on a residual attention mechanism. *Appl. Intell.* **53**(3), 3273–3290 (2023).
32. Huang, J. D. *et al.* FVT: Finger vein transformer for authentication. *IEEE Trans. Instrum. Meas.* **71**, 1–3 (2022).
33. Esmaili, V., Feghhi, M. M. & Shahdi, S. O. Spotting micro-movements in image sequence by introducing intelligent cubic-LBP. *IET Image Process.* **16**(4), 3814–3930 (2022).

### Author contributions

H.Z. provided the idea of algorithm. H.Z. and W.S. carried out the simulations, arranged the architecture, and drafted the manuscript. L.L. supervised the work and revised the manuscript. Both authors read and approved the final manuscript. All authors contributed to the article and approved the submitted version.

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to L.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024